

with z^2 . We can express this similarity with an appropriate holomorphic change of variables. Indeed, suppose that $z = w + 1/w$. Then when w changes to w^2 , z changes to $w^2 + 1/w^2$. But this equals

$$(w + 1/w)^2 - 2 = z^2 - 2 = Q_{-2}(z).$$

The reason this does not show that Q_0 and Q_{-2} are equivalent is that the change of variables cannot be inverted. However, in a suitable region it can. If $z = w + 1/w$, then $w^2 - wz + 1 = 0$. Solving this quadratic equation we find that $w = \frac{1}{2}(z \pm \sqrt{z^2 - 4})$, which leaves us with the problem of which square root to take. It can be shown that for one choice $|w| < 1$ and for the other choice $|w| > 1$, as long as z does not lie in the interval $[-2, 2]$. If we always choose the square root for which $|w| > 1$, then it turns out that the resulting function of z is a continuous function (in fact, holomorphic) from the set $\mathbb{C} \setminus [-2, 2]$ of complex numbers not in $[-2, 2]$ to the set $\{w : |w| > 1\}$ of complex numbers of modulus greater than 1.

Once this is established, it follows that the behavior of Q_{-2} on the set $\mathbb{C} \setminus [-2, 2]$ is topologically the same as the behavior of Q_0 on the set $\{w : |w| > 1\}$. In particular, points outside $\mathbb{C} \setminus [-2, 2]$ have orbits that tend to infinity under iteration by Q_{-2} . Therefore, the attracting basin $A_{-2}(\infty)$ of Q_{-2} is $\mathbb{C} \setminus [-2, 2]$, and the filled Julia set K_{-2} and the Julia set J_{-2} are both equal to $[-2, 2]$.

Let us write $\psi_{-2}(w)$ for $w + 1/w$. The function ψ_{-2} , which we used to change variables, maps circles of radius greater than 1 onto ellipses, and takes radial lines $R_0(\theta)$ that consists of all complex numbers of some given argument θ and modulus greater than 1 to half-branches of hyperbolas. Since the ratio of $\psi_{-2}(w)$ to w tends to 1 as $w \rightarrow \infty$, each radial line will be the asymptote of the corresponding hyperbola half-branch (see figure 5).

It turns out that what we have just done for the polynomial Q_{-2} can be done for any quadratic polynomial Q_c . That is, for sufficiently large complex numbers there is a holomorphic function, denoted φ_c , called the *Böttcher map*, that changes variables in such a way that Q_c turns into Q_0 , in the sense that $\varphi_c(Q_c(z)) = \varphi_c(z)^2$. (The map ψ_{-2} described above is the *inverse* of the Böttcher map in the case $c = -2$, rather than the map itself.) After the change of variables, the new coordinates are called *Böttcher coordinates*.

More generally, for all monic polynomials P (i.e., polynomials with leading coefficient 1) there is a unique holomorphic change of variables φ_P that converts P

into the function $z \mapsto z^d$ for large enough z , in the sense that $\varphi_P(P(z)) = \varphi_P(z)^d$, and has the property that $(\varphi_P(z)/z) \rightarrow 1$ as $z \rightarrow \infty$. The inverse of φ_P is written ψ_P .

2.7.2 Potentials

As we have noted already, if one repeatedly squares a complex number z of modulus greater than 1, then it will escape to infinity. The larger the modulus of z , the faster the iterates will tend to infinity. If instead of squaring, one applies a monic polynomial P of degree d , then for large enough z it is again true that the iterates $z, P(z), P^2(z), \dots$ tend to infinity. It follows from the formula $\varphi_P(P(z)) = \varphi_P(z)^d$ that $\varphi_P(P^k(z)) = \varphi_P(z)^{d^k}$. Therefore, the speed at which the iterates tend to infinity depends not on $|z|$ but on $|\varphi_P(z)|$: the larger the value of $|\varphi_P(z)|$, the faster the convergence. For this reason, the level sets of $|\varphi_P|$, that is, sets of the form $\{z \in \mathbb{C} : |\varphi_P(z)| = r\}$, are important.

For many purposes it is useful to look not at the function φ_P itself but at the function $g_P(z) = \log |\varphi_P(z)|$. This function is called the *potential*, or *Green's function*. It has the same level sets as $|\varphi_P(z)|$, but has the advantage that it is a HARMONIC FUNCTION [IV.24 §5.1].

Clearly, g_P is defined whenever φ_P is defined. But we can in fact extend the definition of g_P to the whole of the attracting basin $A_P(\infty)$. Given any z for which the iterates $P^k(z)$ tend to infinity, one chooses some k such that $\varphi_P(P^k(z))$ is defined and one sets $g_P(z)$ to be $d^{-k} \log |\varphi_P(P^k(z))|$. Notice that $\varphi_P(P^{k+1}(z)) = \varphi_P(P^k(z))^d$, so $\log |\varphi_P(P^{k+1}(z))| = d \log |\varphi_P(P^k(z))|$, from which it is easy to deduce that the value of $d^{-k} \log |\varphi_P(P^k(z))|$ does not depend on the choice of k .

The level sets of g_P are called *equipotentials*. Notice that the equipotential of potential $g_P(z)$ is mapped by P onto the equipotential of potential $g_P(P(z)) = d g_P(z)$. As we shall see, useful information about the dynamics of the polynomial P can be deduced from information about its equipotentials.

If ψ_P is defined everywhere on the circle C_r of radius r , for some $r > 1$, then it maps it to $\{z : |\varphi_P(z)| = r\}$, which is the equipotential of potential $\log r$. For large enough r , this equipotential is a simple closed curve encircling K_P , and it shrinks as r decreases. It is possible for two parts of this curve to come together so that it forms a figure-of-eight shape and then splits into two, like an amoeba dividing, but

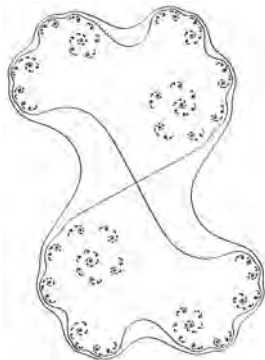


Figure 6 The Julia set of a quadratic polynomial Q_c for which the critical point 0 escapes to infinity under iteration. The Julia set is totally disconnected. The figure-of-eight-shaped curve with 0 at its intersection point is the equipotential through 0. The simple closed curve surrounding it is the equipotential through the critical value c .

this can happen only if the curve crosses a critical point of P . Therefore, if all the critical points of P belong to the filled Julia set K_P (as in the example Q_{-2} , where $0 \in K_{-2} = [-2, 2]$), then it cannot happen. In this case, the Böttcher map φ_P can be defined on the whole of the attracting basin $A_P(\infty)$, and it is a bijection from $A_P(\infty)$ to the attracting basin $A_0(\infty) = \{w \in \mathbb{C} : |w| > 1\}$ of the polynomial z^d . There are equipotentials of potential t for every $t > 0$ and they are all simple closed curves. (Compare with figure 5.) As t approaches 0, the equipotential of potential t , together with its interior, forms a shape that gets closer and closer to the filled Julia set K_P . It follows that K_P is a connected set, as is the Julia set J_P .

On the other hand, if at least one of the critical points in the plane belongs to $A_P(\infty)$, then at a certain point the image of C_r splits into two or more pieces. In particular, the equipotential containing the fastest escaping critical point (i.e., the critical point with the highest value of the potential g_P) has at least two loops, as is illustrated in figure 6. The inside of each loop is mapped by P onto the inside of the equipotential of the corresponding critical value, which is a simple closed curve (since the potential of the critical value is greater than the potential of any critical point). Inside each loop there must be points from the filled Julia set K_P , so this set must be disconnected. The Böttcher map can always be defined on the outside of the equipotential of the fastest escaping critical point and can therefore always be applied to the fastest escaping critical value.

If Q_c is a quadratic polynomial for which 0 escapes to infinity under iteration, then the filled Julia set turns out to be *totally disconnected*, which means that the connected components of K_c are points. None of these points is isolated: they can all be obtained as limits of sequences of other points of K_c . A set which is compact, totally disconnected, and with no isolated points is called a CANTOR SET [III.17], since such a set is homeomorphic to Cantor's middle-thirds set. Note that in this case $K_c = J_c$. For Q_c we have the following dichotomy: the Julia set J_c is connected if 0 has a bounded orbit, and it is totally disconnected if 0 escapes to infinity under iteration. We shall return to this dichotomy when we come to define the Mandelbrot set later in this article.

2.7.3 External Rays of Polynomials with Connected Julia Set

We have just obtained information by looking at the images under ψ_P of circles of radius greater than 1. We can obtain complementary information from the images of *radial lines*, which cut all these circles at right angles. If the Julia set is connected, then, as we saw in the discussion of potentials, the Böttcher map φ_P is a bijection from the attracting basin $A_P(\infty)$ to the attracting basin of z^d , which is the complement $\{w : |w| > 1\}$ of the closed unit disk. As before, let $R_0(\theta)$ denote the half-line that consists of all complex numbers of argument θ and modulus greater than 1. Because $(\varphi_P(z)/z) \rightarrow 1$ as $z \rightarrow \infty$, the image of $R_0(\theta)$ under ψ_P is a half-infinite curve consisting of points with arguments getting closer and closer to θ . This curve is denoted by $R_P(\theta)$, and is known as the *external ray of argument θ* of P . Note that $R_0(\theta)$ is the external ray of argument θ of z^d .

One can think of equipotentials as contour lines of the potential function, and of external rays as the lines of steepest ascent. Between the two of them, they provide a parametrization of the attracting basin, just as modulus and argument provide a parametrization of $\{z : |z| > 1\}$: if you know the potential at a certain complex number z , and you also know which external ray it lies on, then you know what z is. Moreover, a ray of argument θ is mapped by P onto the ray of argument $d\theta$, just as, when a number z lies on the half-line $R_0(\theta)$, then z^d lies on the half-line $R_0(d\theta)$.

We say that an external ray *lands* if $\psi_P(re^{2\pi i\theta})$ converges to a limit as $r \searrow 1$. If this happens, then the limit is called the *landing point*. However, it may happen that

the end of the ray oscillates so much that there is a continuum of different limit points. In this case the ray is *nonlanding*. It can be shown that all rational rays land. Since a rational ray is either periodic or pre-periodic under iteration by P , the landing point of a rational ray must be either a periodic or a pre-periodic point in the Julia set. Much of the structure of the Julia set can be picked up from knowledge about common landing points. In the example illustrated in figure 2, the closures of the three Fatou components containing the critical orbit have one point in common. This point is a repelling fixed point and the common landing point of the rays of argument $\frac{1}{7}$, $\frac{2}{7}$, $\frac{4}{7}$. The rays of argument $\frac{1}{7}$ and $\frac{2}{7}$ are adjacent to the Fatou component containing the critical value c_0 . These two arguments will show up again in the parameter plane and tell us where c_0 is situated.

2.7.4 Local Connectedness

In the example illustrated in figure 5 the inverse of the Böttcher map (the function ψ_{-2}) is defined on the set $\{w : |w| > 1\}$ of all complex numbers w of modulus greater than 1. However, it can be continuously extended to a function defined on the larger set $\{w : |w| \geq 1\}$. If we use the formula $\psi_{-2}(w) = w + 1/w$, then we have $\psi_{-2}(e^{2\pi i\theta}) = 2\cos(2\pi\theta)$, which is the landing point of the external ray $R_{-2}(\theta)$. For an arbitrary connected filled Julia set K_P , we have the following result of Carathéodory: the inverse ψ_P of the Böttcher map has a continuous extension from $\{w : |w| > 1\}$ to $\{w : |w| \geq 1\}$ if and only if K_P is *locally connected*. To understand what this means, imagine a set that is shaped like a comb. From any point in this set to any other point there is a continuous path that lies in the set, but it is possible for the two points to be very close and for the shortest path to be very long. This happens, for example, if the two points are the ends of neighboring teeth of the comb. A connected set X is called locally connected if every point has arbitrarily small connected neighborhoods. It is possible to build comb-like sets (with infinitely many teeth) that contain points for which all connected neighborhoods have to be large. The filled Julia sets in the examples in figures 2–5 are locally connected, but there are examples of filled Julia sets that are not locally connected. When K_P is locally connected, then all external rays land, and the landing point is a continuous function of the argument. Under these circumstances, we have a natural and useful parametrization of the Julia set J_P .

2.8 The Mandelbrot Set M

We shall now restrict our attention to quadratic polynomials of the form Q_c . These are parametrized by the complex number c , and in this context we shall refer to the complex plane as the *parameter plane*, or *c-plane*. We would like to understand the family of dynamical systems that arise when we iterate the polynomials Q_c . Our goal will be to do this by dividing the c -plane into regions that correspond to polynomials with qualitatively the same dynamics. These regions will be separated by their boundaries, which together form the so-called *bifurcation set*. This consists of “unstable” c -values: that is, values of c for which there are other values arbitrarily nearby that give rise to qualitatively different dynamical behavior. In other words, a parameter c belongs to the bifurcation set if a small perturbation of c can make an important difference to the dynamics.

Recall the dichotomy that we stated earlier: the Julia set J_c is connected if the critical point 0 belongs to the filled Julia set K_c and is totally disconnected if 0 belongs to the attracting basin $A_c(\infty)$. This dichotomy motivates the following definition: the *Mandelbrot set* M consists of the c -values for which J_c is connected. That is,

$$M = \{c \in \mathbb{C} \mid Q_c^k(0) \not\rightarrow \infty \text{ as } k \rightarrow \infty\}.$$

Since the Julia set represents the chaotic part of the dynamical system given by Q_c , the dynamical behavior is certainly qualitatively affected by whether c belongs to M or not. We have therefore made a start toward our goal, but the division of the plane into M and $\mathbb{C} \setminus M$ is very coarse, and it does not obviously give us the complete understanding we are looking for.

The important set is in fact not M , but its boundary ∂M , which is illustrated in figure 7. Notice that this set has a number of “holes” (in fact, infinitely many). The Mandelbrot set itself is obtained by filling in all these holes. More precisely, the complement of ∂M consists of an infinite collection of connected components, of which one, the outside of the set, stretches off to infinity, while all the others are bounded. The “holes” are the bounded components.

This definition is similar to the definition of the Julia set of a polynomial. It is easy to define the filled Julia set, and the Julia set is then defined as its boundary. The Julia set provides a lot of structure in the dynamical plane, the z -plane. The Mandelbrot set is similarly easy to define, and its boundary provides a lot of structure in the c -plane. Remarkably, even though each Julia

PUP: Tim confirms that the figure does indeed show what is described here.

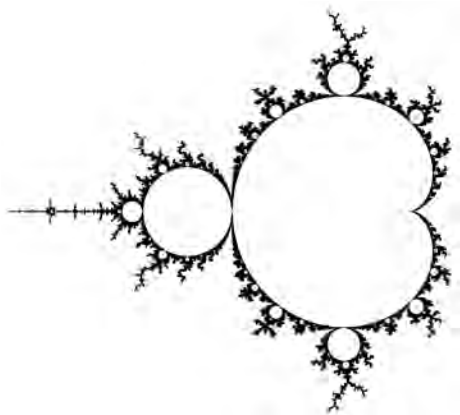


Figure 7 The boundary ∂M of the Mandelbrot set.

set concerns just one dynamical system, while the Mandelbrot set concerns an entire family of systems, there are close analogies between them, as will become clear.

Pioneering work on holomorphic dynamics in general and quadratic polynomials in particular was carried out in the early 1980s by Adrien Douady and John H. Hubbard. They introduced the name “Mandelbrot set” and proved several results about it. In particular, they defined a sort of Böttcher map, denoted by Φ_M , for the Mandelbrot set, which is a map from the complement of the Mandelbrot set to the complement of the closed unit disk.

The definition of Φ_M is actually quite simple: for each c let $\Phi_M(c)$ equal $\varphi_c(c)$, where φ_c is the Böttcher map for the parameter c . However, Douady and Hubbard did more than merely define Φ_M : they proved that it is a holomorphic bijection with holomorphic inverse.

Once we have Φ_M we can make further definitions, just as we did with the Böttcher map. For instance, we can define a *potential* G on the complement of the Mandelbrot set by setting $G(c) = g_c(c) = \log |\Phi_M(c)|$. An *equipotential* is then a level set of Φ_M (that is, a set of the form $\{c \in \mathbb{C} : |\Phi_M(c)| = r\}$ for some $r > 1$) and the *external ray of argument* θ is the set $\{c \in \mathbb{C} : \arg(\Phi_M(c)) = 2\pi\theta\}$ (that is, the inverse image of a radial line $\mathcal{R}_0(\theta)$). The latter is denoted by $\mathcal{R}_M(\theta)$ and it is asymptotic to the radial line of argument θ . The rational external rays are known to land (see figure 8).

It follows from the above that as t approaches zero, the equipotential of potential t , together with its interior, gets closer and closer to M : that is, M is the intersection of all such sets. Hence, M is a connected, closed, bounded subset of the plane.

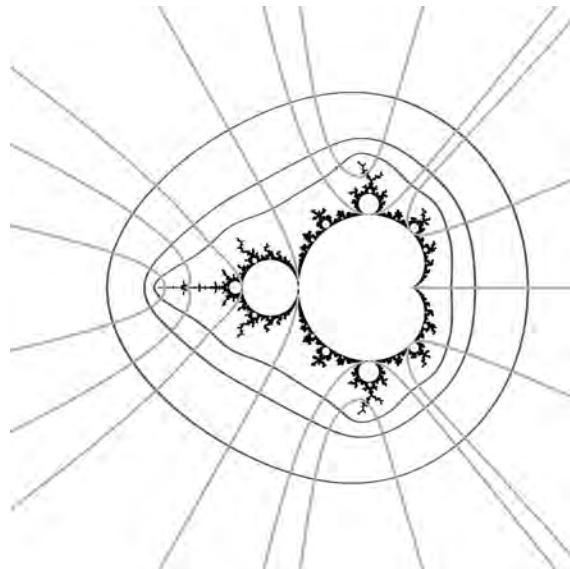


Figure 8 Some equipotentials of M and the external rays of arguments θ of periods 1, 2, 3, and 4. In counterclockwise direction the arguments between 0 and $\frac{1}{2}$ are $0, \frac{1}{15}, \frac{2}{15}, \frac{1}{7}, \frac{3}{15}, \frac{4}{15}, \frac{2}{7}, \frac{1}{3}, \frac{6}{15}, \frac{3}{7},$ and $\frac{7}{15}$; and symmetrically in clockwise direction they are $1 - \theta$ with θ as above. The external rays of argument $\frac{1}{2}$ and $\frac{2}{7}$ are landing at the root point of the hyperbolic component that has c_0 , the parameter value of the Douady rabbit in figure 2, as its center. The rays of argument $\frac{3}{15}$ and $\frac{4}{15}$ are landing at the root point of the copy of M shown in figure 9.

2.8.1 J -Stability

As we have mentioned and as figure 7 suggests, the complement of ∂M has infinitely many connected components. These components are of great dynamical significance: if c and c' are two parameters taken from the same component, then the dynamical systems arising from Q_c and $Q_{c'}$ can be shown to be essentially the same. To be precise, they are *J-equivalent*, which means that there is a continuous change of variables that converts the dynamics on one Julia set to the dynamics on the other. If c belongs to the boundary ∂M , then there are parameter values c' arbitrarily close to c for which Q_c and $Q_{c'}$ are not *J-equivalent*, so ∂M is the “bifurcation set with respect to *J*-stability.” We shall comment on the global structural stability later.

2.8.2 Hyperbolic Components

From now on, we shall use the word “component” to refer to the holes of the Mandelbrot set—that is, to the bounded components of the complement of ∂M .

PUP: Tim says that the caption adequately explains what's going on and that all is fine.

We start by considering the component containing $c = 0$, the central component \mathcal{H}_0 . Recall from section 2.3 that, after a suitable change of variables, one can change the polynomial $F_\lambda(z) = \lambda z + z^2$ into the polynomial Q_c , where the parameters λ and c are related by the equation $c = \frac{1}{2}\lambda - \frac{1}{4}\lambda^2$. The parameter λ has a dynamical meaning: the origin is a fixed point of F_λ and λ is its multiplier. This knowledge tells us that the corresponding Q_c has a fixed point of multiplier λ ; we denote the fixed point by α_c . For $|\lambda| < 1$ the fixed point is attracting.

The unit disk $\{\lambda : |\lambda| < 1\}$ corresponds to the central component \mathcal{H}_0 , and the function that takes a parameter c in \mathcal{H}_0 to the corresponding parameter λ in the unit disk is called the *multiplier map*, and is denoted by $\rho_{\mathcal{H}_0}$. Thus, $\rho_{\mathcal{H}_0}(c)$ is the multiplier of the fixed point α_c of the polynomial Q_c . The multiplier map $\rho_{\mathcal{H}_0}$ is a holomorphic isomorphism from \mathcal{H}_0 to the unit disk. As we have just seen, the inverse map is given by $\rho_{\mathcal{H}_0}^{-1}(\lambda) = \frac{1}{2}\lambda - \frac{1}{4}\lambda^2$. This map extends continuously to the unit circle, and thereby gives us a parametrization of the boundary of the central component \mathcal{H}_0 by points λ of modulus 1. The image of the unit circle under the map $\lambda \mapsto \frac{1}{2}\lambda - \frac{1}{4}\lambda^2$ is a *cardioid*. This explains the heart-like shape of the largest part of the Mandelbrot set, which can be seen in figure 7.

Any quadratic polynomial has two fixed points counted with multiplicity (in fact, two distinct ones unless $c = \frac{1}{4}$). The central component \mathcal{H}_0 is characterized as the component of c -values for which Q_c has an attracting fixed point. For any c outside the cardioid, Q_c has two repelling fixed points, but it may have an attracting periodic orbit of a period greater than 1. It is an important fact that the attracting basin of an attracting periodic orbit always contains a critical orbit. Therefore, for any quadratic polynomial there can be at most one attracting periodic orbit.

We call a component \mathcal{H} of the Mandelbrot set a *hyperbolic component* if, for every parameter c in \mathcal{H} , the polynomial Q_c has an attracting periodic orbit. For any given hyperbolic component, the periods of the attracting periodic orbits will be the same. There is a corresponding multiplier map $\rho_{\mathcal{H}}$, from \mathcal{H} to the unit disk, which assigns to each parameter c in \mathcal{H} the multiplier of the attracting periodic orbit. This multiplier map is always a holomorphic isomorphism that extends continuously to the boundary $\partial\mathcal{H}$ of \mathcal{H} .

The points $\rho_{\mathcal{H}}^{-1}(0)$ and $\rho_{\mathcal{H}}^{-1}(1)$ are called the *center* and the *root* of \mathcal{H} . The center of \mathcal{H} is the unique c in \mathcal{H} for which the periodic orbit of Q_c is super-attracting.

As for the root, if the period of the component is k , then it will be the landing point for a pair of external rays of periodic arguments of period k . (For the central component \mathcal{H}_0 there is only one ray assigned.) Conversely, every external ray with such an argument lands at the root point of a hyperbolic component of period k . Thus, the arguments of these rays give addresses to the hyperbolic components. This can be seen in figure 8, from which one can read off the mutual positions of all the components of periods 1–4.

As a consequence of the above, the number of hyperbolic components corresponding to a certain period k can be determined both as the number of roots in the polynomial $Q_c^k(0)$ that are not roots in $Q_c^\ell(0)$ for some $\ell < k$ and also as the number of pairs of rational arguments with denominator $2^k - 1$ that cannot be expressed with denominator $2^\ell - 1$ for some $\ell < k$.

For any component \mathcal{H} with center c_0 let $\mathcal{R}_M(\theta_-)$ and $\mathcal{R}_M(\theta_+)$ be the pair of rays landing at the root point. Then, in the dynamical plane of Q_{c_0} , the pair of rays $\mathcal{R}_{c_0}(\theta_-)$ and $\mathcal{R}_{c_0}(\theta_+)$ are adjacent to the Fatou component of Q_{c_0} containing c_0 , and they land at the root point of that Fatou component.

2.8.3 Structural Stability

Suppose that Q_c has a super-attracting periodic orbit of period k , and let z_0 be a point in this orbit. Then $Q_c^k(z_0) = z_0$, and the derivative of Q_c^k at z_0 is 0. It follows from the chain rule that there is at least one z_i in the orbit at which the derivative of Q_c is 0: that is, 0 belongs to the orbit. Therefore, the center of a hyperbolic component cannot be structurally stable, since the critical orbit of the center-polynomial is finite, but it is infinite for all nearby polynomials. However, if we remove from the complex plane not just ∂M but also all the centers of hyperbolic components, then we obtain the splitting we have been looking for: any connected component of the remaining set forms a structurally stable region. For any pair of parameter values c and c' in such a component, Q_c and $Q_{c'}$ are conjugate, meaning that there is a continuous change of variables in the plane that converts the dynamics of one polynomial into those of the other.

2.8.4 Conjectures

The above discussion raises an obvious question: we have a good understanding of the hyperbolic components of the complement of ∂M , but are there compo-

nents that are *not* hyperbolic? The following conjecture expresses a widely held belief, but it is as yet unproved.

The hyperbolicity conjecture. *All the bounded components of the complement of ∂M are hyperbolic.*

The hyperbolicity conjecture can be stated in greater generality for rational functions, where it says that every rational function can be approximated arbitrarily closely by a *hyperbolic rational function*. Here, “hyperbolic” means that the dynamics is expanding on the Julia set. We shall not go further into this, but only mention that the dynamics on the Julia set is expanding for every Q_c with c in a hyperbolic component of M , and also in the unbounded component, the complement of M . The Julia set J_c can in these cases be thought of as a “strange repeller”: the dynamics is chaotic and the geometry is fractal (except for $c = 0$).

The main conjecture about the Mandelbrot set is, however, the following.

The local connectivity conjecture. *The Mandelbrot set is locally connected.*

This conjecture, often referred to as MLC, is important for many reasons. To begin with, it is known that it implies the hyperbolicity conjecture. Second, if M is locally connected, then Ψ_M , the inverse of Φ_M , which is a holomorphic bijection from the set outside the closed unit disk to the complement of the Mandelbrot set, has a continuous extension to the unit circle, and all external rays land in a continuous manner. This would give us a useful parametrization of ∂M . One can then give a beautifully simple abstract combinatorial description of M , despite the fact that ∂M is a complicated fractal. (Mitsuhiro Shishikura has proved that the HAUSDORFF DIMENSION [III.17] of ∂M is the maximum possible in the plane, namely 2.)

2.9 Universality of M

The Mandelbrot set is remarkably ubiquitous. For example, homeomorphic copies of M appear inside M itself, as is apparent from figure 9. Inside other families of holomorphic mappings that depend holomorphically on some parameter, we again find homeomorphic copies of M . For this reason, M is said to be *universal*. Douady and Hubbard have captured the reason behind the phenomenon of universality by defining a notion of a *quadratic-like mapping*. The k th iterate of a quadratic polynomial is globally a polynomial

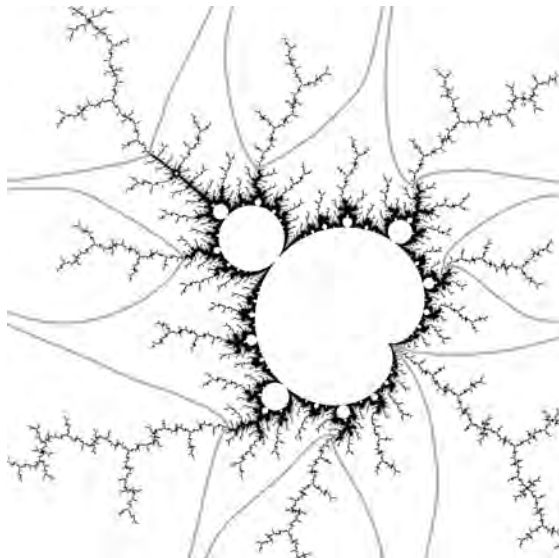


Figure 9 A copy of M within M . The address of the copy is given by the arguments of the two external rays that land at the cusp, the root point of the copy. Here the arguments are $\frac{3}{15}$ and $\frac{4}{15}$. Compare with figure 8. The rays are drawn to indicate where the “decorations” should be cut off in order to have the bare copy of M .

of degree 2^k , but locally it may behave like a quadratic polynomial. The same is true for a rational function or an iterate of it. By a quadratic-like mapping we mean a triple (f, V, W) where V and W are open simply connected domains (that is, connected open sets without holes), $\bar{V} \subset W$, and f is a holomorphic map that maps V onto W with degree 2. (This means that every point in W has two preimages, up to multiplicity, in V .) Such a map f has a single critical point ω in V , and behaves in many ways like a quadratic polynomial. The filled Julia set K_f is defined as the set of points z in V for which the iterates $f^k(z)$ stay in V for all $k \geq 0$. A dichotomy similar to the one for quadratic polynomials holds for quadratic-like mappings as well: K_f is connected if and only if the critical point ω is contained in K_f . For any quadratic-like mapping with a connected filled Julia set, Douady and Hubbard have defined a strategy, called *straightening*, which associates with the mapping a unique c -value in M . For a family of quadratic-like mappings $\{f_\lambda\}_{\lambda \in \Lambda}$ the Mandelbrot set M_Λ is defined as the set of λ for which K_{f_λ} is connected. We obtain through straightening a mapping $\Xi : M_\Lambda \rightarrow M$, which takes λ to the uniquely associated c -value.

In the copy of M shown in figure 9, the “center” associated with $c = 0$ in M corresponds to a polynomial Q_{c_0} for which the critical point 0 is periodic of period 4, and for which a suitable restriction of the fourth iterate $f_{c_0} = Q_{c_0}^4$ is quadratic-like from V_0 to its image W_0 . Moreover, there is a neighborhood \mathcal{V}_0 of c_0 in the c -plane such that for any c in \mathcal{V}_0 the restriction of $f_c = Q_c^4$ to V_0 is a quadratic-like map from V_0 to its image W_c , and such that the map \mathcal{E} is a homeomorphism from $M_{\mathcal{V}_0}$ to M .

The infinitely many copies of M that appear inside M may suggest that M has a self-similarity property. However, there is another phenomenon that pulls in the opposite direction. The c -values for which the critical point 0 is pre-periodic form a dense subset of ∂M . If \tilde{c} is one of these special c -values, then there are two contexts in which one may look at magnifications of smaller and smaller neighborhoods of \tilde{c} : the first is the Julia set $J_{\tilde{c}}$ of the polynomial $Q_{\tilde{c}}$ in neighborhoods of $z = \tilde{c}$, and the second is the Mandelbrot set in neighborhoods of $c = \tilde{c}$. It turns out that the pictures are *asymptotically similar*, which means that the greater the magnification, and the smaller the neighborhood, the more similar the two pictures become.

This is an extraordinary fact. Indeed, it may even seem to be impossible, since in any neighborhood of \tilde{c} the Mandelbrot set contains infinitely many copies of itself, while the Julia set is known to contain no such copies. The explanation for the apparent paradox is that the copies of the Mandelbrot set get smaller very quickly as their distance to \tilde{c} decreases. Hence, if one magnifies a small enough neighborhood, the copies that are there are practically invisible.

2.10 Newton's Method Revisited

Let us return briefly to Newton's method for polynomials. Consider any polynomial P of degree $d \geq 2$ that has only simple roots. Then the Newton function N_P is a rational function of degree d , and each simple root of P is a super-attracting fixed point of N_P . For quadratic polynomials the number of roots of P coincides with the number of critical points of N_P (since $2d - 2 = 2$ when $d = 2$). For polynomials of degree $d > 2$ there are more critical points than the roots can account for.

Cayley considered Newton's method for quadratic polynomials with two distinct roots $P(z) = (z - r_1)(z - r_2)$. He showed that the function $\mu(z) = (z - r_1)/(z - r_2)$, which maps the root r_1 onto 0 and the root r_2 onto ∞ , provides a change of variables that turns

N_P into the quadratic polynomial Q_0 on the Riemann sphere $\hat{\mathbb{C}}$. When one translates the dynamics of Q_0 to the dynamics of Newton's method one finds that the unit circle corresponds to the bisector of r_1 and r_2 and that all points in the half-plane containing r_i , $i = 1, 2$, are therefore attracted to r_i under iteration by N_P .

Cayley announced that he would write about Newton's iteration for cubic polynomials. However, it took about a hundred years before any such paper appeared. For a cubic polynomial P with three simple roots the Newton function N_P has three super-attracting fixed points, each of which gives rise to an attracting basin. The Julia set of N_P is the common boundary of these three basins, and is therefore a complicated fractal set. Moreover, N_P has an extra critical point since $2d - 2 = 4$ for $d = 3$. The extra critical point may be attracted to one of the roots under iteration, or it can have its own independent behavior. In order to catch the behavior of all cubic polynomials under Newton's iteration (except the one with one root of multiplicity three) it is sufficient to consider the one-parameter family of polynomials $P_\lambda(z) = (z - 1)(z - \frac{1}{2} - \lambda)(z - \frac{1}{2} + \lambda)$. The extra critical point for the corresponding Newton function N_λ then turns out to be at the origin. Suppose that we associate three colors, for instance red, blue, and green, with the three roots $1, \frac{1}{2} + \lambda, \frac{1}{2} - \lambda$. We can then color the λ -plane, which is the parameter plane in this context, as follows. A parameter value λ is colored red, blue, or green if the critical point 0 is attracted under iteration by N_λ to the root of that color. If it is not attracted to any of the three roots, then we color with a fourth color, yellow, say. The universality of the Mandelbrot set is thereby demonstrated: in the λ -plane one can observe yellow copies of it, which one can explain by showing that families of suitably restricted iterates of N_λ are quadratic-like.

3 Concluding Remarks

We have illustrated several results in holomorphic dynamics through examples, including the transferring of definitions and results from the dynamical planes to the parameter plane. The structures of the filled Julia sets and the Mandelbrot set are partly understood through analysis of their complements, linked together via the Böttcher maps φ_c and Φ_M . The functions that are used for changing variables in J -stability and structural stability are examples of so-called *quasi-conformal mappings*. This is a concept that was introduced into holomorphic dynamics in the early 1980s by

Dennis Sullivan. They are indispensable for discussing change of *complex structure*, *straightening*, *holomorphic motion*, *surgery*, and many other phenomena. The interested reader is referred to the books listed below. The first two contain expository papers, the third is a graduate textbook, and the fourth is a collection of papers. They all contain many further references.

Acknowledgments. The computer drawings in this article were obtained from a program written by Christian Henriksen.

Further Reading

- Devaney, R. L., and L. Keen, eds. 1989. *Chaos and Fractals. The Mathematics Behind the Computer Graphics*. Proceedings of Symposia in Applied Mathematics, volume 39. Providence, RI: American Mathematical Society.
- . 1994. *Complex Dynamical Systems. The Mathematics Behind the Mandelbrot and Julia Sets*. Proceedings of Symposia in Applied Mathematics, volume 49. Providence, RI: American Mathematical Society.
- Lei, T., ed. 2000. *The Mandelbrot Set, Theme and Variations*. London Mathematical Society Lecture Note Series, volume 274. Cambridge: Cambridge University Press.
- Milnor, J. 1999. *Dynamics in One Complex Variable*. Weisbaden: Vieweg.

IV.15 Operator Algebras

Nigel Higson and John Roe

1 The Beginnings of Operator Theory

We can ask two basic questions about any equation, or system of equations: is there a solution, and, if there is, is it unique? Experience with finite systems of linear equations indicates that the two questions are interconnected. Consider for instance the equations

$$\begin{aligned} 2x + 3y - 5z &= a, \\ x - 2y + z &= b, \\ 3x + y - 4z &= c. \end{aligned}$$

Notice that the left-hand side of the third equation is the sum of the left-hand sides of the first two. As a result, no solution to the system exists unless $a + b = c$. But if $a + b = c$, then any solution of the first two equations is also a solution of the third; and in any linear system involving more unknowns than equations, solutions, when they exist, are never unique. In the present case, if (x, y, z) is a solution, then so is $(x+t, y+t, z+t)$, for any t . Thus the same phenomenon (a linear relation among the equations) that prevents

the system from admitting solutions in some cases also prevents solutions from being unique in other cases.

To make the relation between existence and uniqueness of solutions more precise, consider a general system of linear equations of the form

$$\begin{aligned} k_{11}u_1 + k_{12}u_2 + \cdots + k_{1n}u_n &= f_1, \\ k_{21}u_1 + k_{22}u_2 + \cdots + k_{2n}u_n &= f_2, \\ &\vdots \\ k_{n1}u_1 + k_{n2}u_2 + \cdots + k_{nn}u_n &= f_n \end{aligned}$$

consisting of n equations in n unknowns. The scalars k_{ji} form a matrix of coefficients and the problem is to solve for the u_i in terms of the f_j . The general theorem illustrated by our particular numerical example above is that the number of linear conditions that the f_j must satisfy if a solution is to exist is equal to the number of arbitrary constants appearing in the general solution when a solution does exist. To use a more technical vocabulary, the dimension of the **KERNEL** [I.3 §4.1] of the matrix $K = \{k_{ji}\}$ is equal to the dimension of its cokernel. In the example, these numbers are both 1.

A little more than a hundred years ago, FREDHOLM [VI.66] made a study of *integral equations* of the type

$$u(y) - \int k(y, x)u(x) dx = f(y).$$

These arose from questions in theoretical physics, and the problem was to solve for the function u in terms of the function f . Since an integral can be thought of as a limit of finite sums, Fredholm's equation is an infinite-dimensional counterpart of the finite-dimensional linear systems considered above, in which vectors with n components are replaced by functions with values at infinitely many different points x . (Strictly speaking, Fredholm's equation is analogous to a matrix equation of the type $u - Ku = f$ rather than $Ku = f$. The altered form of the left-hand side has no effect on the overall behavior of the matrix equation, but it does considerably alter the behavior of the integral equation. As we shall see, Fredholm was fortunate to work with a class of equations whose behavior mirrors that of matrix equations very closely.)

A very simple example is

$$u(y) - \int_0^1 u(x) dx = f(y).$$

To solve this equation, it helps to observe that the quantity $\int_0^1 u(x) dx$, when thought of as a function of y , is a *constant*. Thus in the homogeneous case ($f \equiv 0$), the only possible solutions for $u(y)$ are the constant

functions. On the other hand, for a general function f , solutions exist if and only if the single linear condition $\int_0^1 f(y) dy = 0$ is satisfied. So in this example the dimension of the kernel and the dimension of the cokernel are both 1. Fredholm set out on a systematic exploration of the analogy between matrix theory and integral equations that this example suggests. He was able to prove that, for equations of his type, the dimensions of the kernel and of the cokernel are always finite and equal.

Fredholm's work sparked the imagination of HILBERT [VI.63], who made a detailed study of the *integral operators* that transform $u(y)$ into $\int k(y, x)u(x) dx$, in the special case where the real-valued function k is *symmetric*, meaning that $k(x, y) = k(y, x)$. The finite-dimensional counterpart of Hilbert's theory is the theory of real symmetric matrices. Now if K is such a matrix, then a standard result from linear algebra asserts that there is an orthonormal basis consisting of EIGENVECTORS [I.3 §4.3] for K , or equivalently that there is a *unitary* matrix U such that $U^{-1}TU$ is diagonal. (*Unitary* means that U is invertible and preserves the lengths of vectors: $\|Uv\| = \|v\|$ for all vectors v .) Hilbert obtained an analogous theory for all symmetric integral operators. He showed that there exist functions $u_1(y), u_2(y), \dots$ and real numbers $\lambda_1, \lambda_2, \dots$ such that

$$\int k(y, x)u_n(x) dx = \lambda_n u_n(y).$$

Thus $u_n(y)$ is an *eigenfunction* for the integral operator, with eigenvalue λ_n .

In most cases it is hard to calculate u_n and λ_n explicitly, but calculation is possible when $k(x, y) = \phi(x - y)$ for some periodic function ϕ . If the range of integration is $[0, 1]$ and the period of ϕ is 1, then the eigenfunctions are $\cos(2k\pi y)$, $k = 0, 1, 2, \dots$, and $\sin(2k\pi y)$, $k = 1, 2, \dots$. In this case, the theory of FOURIER SERIES [III.27] tells us that a general function $f(y)$ on $[0, 1]$ can be expanded as the sum of a series $\sum (a_k \cos 2k\pi y + b_k \sin 2k\pi y)$ of cosines and sines. Hilbert showed that, in general, there is an analogous expansion

$$f(y) = \sum a_n u_n(y)$$

in terms of the eigenfunctions for *any* symmetric integral operator. In other words, the eigenfunctions form a *basis*, just as in the finite-dimensional case. Hilbert's result is now called the *spectral theorem* for symmetric integral operators.

1.1 From Integral Equations to Functional Analysis

Hilbert's theorem led to an explosion of activity, since integral operators arise in many different areas of mathematics (including, for example, the DIRICHLET PROBLEM [IV.12 §1] in partial differential equations and the REPRESENTATION THEORY OF COMPACT GROUPS [IV.9 §3]). It was soon recognized that these operators are best viewed as linear transformations on the HILBERT SPACE [III.37] of all functions $u(y)$ such that $\int |u(y)|^2 dy < \infty$. Such functions are called *square-integrable*, and the collection of all of them is denoted $L^2[0, 1]$.

With the important concept of Hilbert space available, it became convenient to examine a much broader range of operators than the integral operators initially considered by Fredholm and Hilbert. Since Hilbert spaces are VECTOR SPACES [I.3 §2.3] and METRIC SPACES [III.58], it made sense to look first at operators from a Hilbert space to itself that are both linear and continuous: these are usually called *bounded* linear operators. The analogue of the symmetry condition $k(x, y) = k(y, x)$ on integral operators is the condition that a bounded linear operator T be *self-adjoint*, which is to say that $\langle Tu, v \rangle = \langle u, Tv \rangle$ for all vectors u and v in the Hilbert space (the angle brackets denote the inner product). A simple example of a self-adjoint operator is the *multiplication operator* by a real-valued function $m(y)$; this is the operator M defined by the formula $(Mu)(y) = m(y)u(y)$. (The finite-dimensional counterpart to a multiplication operator is a diagonal matrix K , which multiplies the j th component of the vector by the matrix entry k_{jj} .)

PUP: repeated 'j' is fine here.

Hilbert's spectral theorem for symmetric integral operators tells us that every such operator can be given a particularly nice form: with respect to a suitable "basis" of $L^2[0, 1]$, namely a basis of eigenfunctions, it will have an infinite diagonal matrix. Moreover, the basis vectors can be chosen to be orthogonal to each other. For a general self-adjoint operator, this is not true. Consider, for instance, the multiplication operator from $L^2[0, 1]$ to itself that takes each square-integrable function $u(y)$ to the function $yu(y)$. This operator has no EIGENVECTORS [I.3 §4.3], since if λ is an EIGENVALUE [I.3 §4.3], then we need $yu(y) = \lambda u(y)$ for every y , which implies that $u(y) = 0$ for every y not equal to λ , and hence that $\int |u(y)|^2 dy = 0$. However, this example is not particularly worrying, since a multiplication operator of this kind is a sort of continuous analogue of the operator defined by a diagonal matrix.

It turns out that if we enlarge our concept of “diagonal” to include multiplication operators, then all self-adjoint operators are “diagonalizable,” in the sense that, after a suitable “change of basis,” they become multiplication operators.

To make this statement precise, we need the notion of the SPECTRUM [III.88] of an operator T . This is the set of complex numbers λ for which the operator $T - \lambda I$ does not have a bounded inverse (here I is the identity operator on Hilbert space). In finite dimensions the spectrum is precisely the set of eigenvalues, but in infinite dimensions this is not always so. Indeed, whereas every symmetric matrix has at least one eigenvalue, a self-adjoint operator, as we have just seen, need not. As a result of this, the spectral theorem for bounded self-adjoint operators is phrased not in terms of eigenvalues but in terms of the spectrum. One way of formulating it is to state that any self-adjoint operator T is *unitarily equivalent* to a multiplication operator $(Mu)(y) = m(y)u(y)$, where the closure of the range of the function $m(y)$ is the spectrum of T . Just as in the finite-dimensional case, a *unitary* is an invertible operator U that preserves the lengths of vectors. To say that T and M are unitarily equivalent is to say that there is some unitary map U , which we can think of as an analogue of a change-of-basis matrix, such that $T = U^{-1}MU$. This generalizes the statement that any real symmetric matrix is unitarily equivalent to a diagonal matrix with the eigenvalues along the diagonal.

PUP: ‘generalizes’
definitely better
than ‘generalizes
to’ here.

1.2 The Mean Ergodic Theorem

A beautiful application of the spectral theorem was found by VON NEUMANN [VI.91]. Imagine a checkerboard on which are distributed a certain number of checkers. Imagine that for each square there is designated a “successor” square (in such a way that no two squares have the same successor), and that every minute the checkers are rearranged by moving each one to its successor square. Now focus attention on a single square and each minute record with a 1 or 0 whether or not there is a piece on the square. This produces a succession of readings R_1, R_2, R_3, \dots like this:

00100110010110100100 \dots

We might expect that over time, the average number of positive readings $R_j = 1$ will converge to the number of pieces on the board divided by the number of squares. If the rearrangement rule is not complicated enough, then this will not happen. For example, in the most extreme case, if the rule designates each square

as its own successor, then the readout will be either 00000 \dots or 11111 \dots , depending on whether or not we chose a square with a piece on it to begin with. But if the rule is sufficiently complicated, then the “time average” $(1/n) \sum_{j=1}^n R_j$ will indeed converge to the number of pieces on the board divided by the number of squares, as expected.

The checkerboard example is elementary, since in fact the only “sufficiently complicated” rules in this finite case are cyclic permutations of the squares of the board, and thus *all* the squares move past our observation post in succession. However, there are related examples where one observes only a small fraction of the data. For instance, replace the set of squares on a checkerboard with the set of points on a circle, and in place of the checkers, imagine that a subset S of a circle is marked as occupied. Let the rearrangement rule be the rotation of points on the circle through some irrational number of degrees. Stationed at a point x of the circle, we record whether x belongs to S , the first rotated copy of S , the second rotated copy of S , and so on to obtain a sequence of 0 or 1 readings as before. One can show that (for nearly every x) the time average of our observations will converge to the proportion of the circle occupied by S .

Similar questions about the relationship between time and space averages had arisen in thermodynamics and elsewhere, and the expectation that time and space averages should agree when the rearrangement rule is sufficiently complex became known as the *ergodic hypothesis*.

Von Neumann brought operator theory to bear on this question in the following way. Let H be the Hilbert space of functions on the squares of the checkerboard, or the Hilbert space of square-integrable functions on the circle. The rearrangement rule gives rise to a unitary operator U on H by means of the formula

$$(Uf)(y) = f(\phi^{-1}(y)),$$

where ϕ is the function describing the rearrangement. Von Neumann’s ergodic theorem asserts that if no non-constant function in H is fixed by U (this is one way of saying that the rearrangement rule is “sufficiently complicated”), then, for every function $f \in H$, the limit

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n U^j f$$

exists and is equal to the constant function whose value everywhere is the average value of f . (To apply this to our examples, take $f(x)$ to be the function that is 1 if the point x is occupied and 0 otherwise.)

Von Neumann's theorem can be deduced from a spectral theorem for unitary operators that is analogous to the spectral theorem for self-adjoint operators. Every unitary operator can be reduced to a multiplication operator, not by real-valued functions but by functions whose values are complex numbers of absolute value 1. The key to the proof then becomes a statement about complex numbers of absolute value 1: if z is such a complex number, different from 1, then the expression $(1/n) \sum_{j=1}^n z^j$ approaches zero as $n \rightarrow \infty$. This in turn is easily proved using the formula for the sum of a geometric series, $\sum_{j=1}^n z^j = z(1 - z^n)/(1 - z)$. (More detail can be found in ERGODIC THEOREMS [V.11].)

1.3 Operators and Quantum Theory

Von Neumann realized that Hilbert spaces and their operators provide the correct mathematical tools to formalize the laws of quantum mechanics, introduced in the 1920s by Heisenberg and Schrödinger.

The *state* of a physical system at any given instant is the list of all the information needed to determine its future behavior. If, for instance, the system consists of a finite number of particles, then classically its state consists of the list of the position and momentum vectors of all the constituent particles. By contrast, in von Neumann's formulation of quantum mechanics one associates with each physical system a Hilbert space H , and a state of the system is represented by a unit vector u in H . (If u and v are unit vectors and v is a scalar multiple of u , then u and v determine the same state.)

Associated with each observable quantity (perhaps the total energy of the system, or the momentum of one particle within the system) is a self-adjoint operator Q on H whose spectrum is the set of all observed values of that quantity (hence the origin of the term "spectrum"). States and observables are related as follows: when a system is in the state described by a unit vector $u \in H$, the *expected value* of the observable quantity corresponding to a given self-adjoint operator Q is the inner product $\langle Qu, u \rangle$. This may not be a value that is ever actually measured: rather, it is the average of values that are obtained from many repeated experiments with the system when it is in the given state u . The relation between states and observables reflects the paradoxical behavior of quantum mechanics: it is possible, and in fact typical, for a system to exist in a "superposed" state, under which repeated identical experiments produce distinct outcomes. A measure-

ment of an observable quantity will produce a determinate outcome if and only if the state of the system is an eigenvector for the operator associated with that quantity.

A distinctive feature of quantum theory is that the operators associated with different observables typically do not commute with one another. If two operators do not commute, then they will typically have no eigenvectors in common, and, as a result, simultaneous measurements of two different observables will typically not result in determinate values for both of them. A famous example is provided by the operators P and Q associated with the position and momentum of a particle moving along a line. They satisfy the *Heisenberg commutation relation*

$$QP - PQ = i\hbar I,$$

where \hbar is a certain physical constant. (This is an instance of a general principle which relates the non-commutativity of observables in quantum mechanics to the *Poisson bracket* of the corresponding observables in classical mechanics: see MIRROR SYMMETRY [IV.16 §§2.1.3, 2.2.1].) As a result, it is impossible for the particle simultaneously to have a determinate momentum and position. This is the *uncertainty principle*.

It turns out that there is an essentially unique way of representing the Heisenberg commutation relation using self-adjoint operators on Hilbert space: the Hilbert space H must be $L^2(\mathbb{R})$; the operator P must be $-i\hbar d/dx$; and the operator Q must be multiplication by x . This theorem allows one to determine explicitly the observable operators for simple physical systems. For example, in a system consisting of a particle on a line subject to a force directed toward the origin which is proportional to the distance from the origin (as if the particle were attached to a spring, anchored at the origin), the operator for total energy is

$$E = -\frac{\hbar^2}{2m} \frac{d^2}{dx^2} + \frac{k}{2} x^2,$$

where k is a constant which determines the overall strength of the force. The spectrum of this operator is the set

$$\{(n + \frac{1}{2})\hbar(k/m)^{1/2} : n = 0, 1, 2, \dots\}.$$

These are therefore the possible values for the total energy of the system. Notice that the energy can assume only a discrete set of values. This is another characteristic and fundamental feature of quantum theory.

Another important example is the operator of total energy for the hydrogen atom. Like the operator above,

this may be realized as a certain explicit partial differential operator. It can be shown that the eigenvalues of this operator form a sequence proportional to $\{-1, -\frac{1}{4}, -\frac{1}{9}, \dots\}$. A hydrogen atom, when disturbed, may release a photon, resulting in a drop in its total energy. The released photon will have energy equal to the difference between the energies of the initial and final states of the atom, and therefore it is proportional to a number of the form $1/n^2 - 1/m^2$. When light from hydrogen is passed through a prism or diffraction grating, bright lines are indeed observed at wavelengths corresponding to these possible energies. Spectral observations of this sort provide experimental confirmation for quantum mechanical predictions.

So far we have discussed states of a quantum system only at a single instant. However, quantum systems evolve in time, just as classical systems do: to describe this evolution we need a law of motion. The time evolution of a quantum system is represented by a family of unitary operators $U_t : H \rightarrow H$, parametrized by the real numbers. If the system is in an initial state u , it will be in the state $U_t u$ after t units of time. Because the passage of s units of time followed by t further units is the same as the passage of $s + t$ units, the unitary operators U_t satisfy the *group law* $U_s U_t = U_{s+t}$. An important theorem of Marshall Stone asserts that there is a one-to-one correspondence between unitary groups $\{U_t\}$ and self-adjoint operators E given by the formula

$$iE = \left(\frac{dU_t}{dt} \right)_{t=0} = \lim_{t \rightarrow 0} \frac{1}{t} (U_t - I).$$

The quantum law of motion is that the *generator* E corresponding in this way to time evolution is the operator associated with the observable “total energy.” When E is realized as a differential operator on a Hilbert space of functions (as in the examples above), this statement becomes a differential equation, the *Schrödinger equation*.

1.4 The GNS Construction

The time-evolution operators U_t of quantum mechanics satisfy the law $U_s U_t = U_{s+t}$. More generally, we define a *unitary representation* of a GROUP [I.3 §2.1] G to be a family of unitary operators U_g , one for each $g \in G$, satisfying the law $U_{g_1 g_2} = U_{g_1} U_{g_2}$ for all $g_1, g_2 \in G$. Originally introduced by FROBENIUS [VI.58] as a tool for the study of finite groups, REPRESENTATION THEORY [IV.9] has become indispensable in mathematics and physics wherever the symmetries of a system must be taken into account.

If U is a unitary representation of G and v is a vector, then $\sigma : g \mapsto \langle U_g v, v \rangle$ is a function defined on G . The law $U_{g_1 g_2} = U_{g_1} U_{g_2}$ implies that σ has an important positivity property, namely

$$\sum_{g_1, g_2 \in G} \overline{a_{g_1}} a_{g_2} \sigma(g_1^{-1} g_2) = \left\| \sum a_g U_g v \right\|^2 \geq 0,$$

for any scalars $a_g \in \mathbb{C}$. A function defined on G and having this positivity property is said to be *positive definite*. Conversely, from a positive-definite function one can build a unitary representation. This *GNS construction* (in honor of Israel Gelfand, Mark Naimark, and Irving Segal) begins by considering the group elements themselves as basis vectors in an abstract vector space. We can attempt to define an inner product on this vector space by means of the formula

$$\langle g_1, g_2 \rangle = \sigma(g_1^{-1} g_2).$$

The resulting object may differ from a genuine Hilbert space in two respects. First, there may be nonzero vectors whose length, as measured by the inner product, is zero (although the hypothesis that σ is positive definite does rule out the possibility that there might be vectors of *negative* length). Second, the COMPLETENESS AXIOM [III.64] of Hilbert space theory may not be satisfied. However, there is a “completion” procedure which fixes both these deficiencies. Applied in the present case, it produces a Hilbert space H_σ that carries a unitary representation of G .

Versions of the GNS construction arise in several areas of mathematics. They have the advantage that the functions on which the constructions are based are easy to manipulate. For instance, convex combinations of positive-definite functions are again positive definite, and this allows geometrical methods to be applied to the study of representations.

1.5 Determinants and Traces

The original works of Fredholm and Hilbert borrowed heavily from traditional concepts of linear algebra, and in particular the theory of DETERMINANTS [III.15]. In view of the complicated definition of the determinant even for finite matrices, it is perhaps not surprising that the infinite-dimensional situation presented extraordinary challenges. Very soon, much simpler alternative approaches were found that avoided determinants altogether. But it is interesting to note that the determinant, or to be more exact the related notion of the trace, has played an important role in recent developments on which we will report later in this article.

The *trace* of an $n \times n$ matrix is the sum of its diagonal entries. As with the determinant, the trace of a matrix A is equal to the trace of BAB^{-1} for any invertible matrix B . In fact, the trace is related to the determinant by the formula $\det(\exp(A)) = \exp(\operatorname{tr}(A))$ (because of the invariance properties of trace and determinant, it is enough to check this for diagonal matrices, where it is easy). In infinite dimensions the trace need not make sense since the sum of the diagonal entries of an $\infty \times \infty$ matrix may not converge. (The trace of the identity operator is a case in point: the diagonal entries are all 1, and if there are infinitely many of them, then their sum is not well-defined.) One way to address this problem is to limit oneself to operators for which the sum is well-defined. An operator T is said to be of *trace class* if, for every two sequences $\{u_j\}$ and $\{v_j\}$ of pairwise orthogonal vectors of length 1, the sum $\sum_{j=1}^{\infty} \langle Tu_j, v_j \rangle$ is absolutely convergent. A trace-class operator T has a well-defined and finite trace, namely the sum $\sum_{j=1}^{\infty} \langle Tu_j, u_j \rangle$ (which is independent of the choice of orthonormal basis $\{u_j\}$).

Integral operators such as those appearing in Fredholm's equation provide natural examples of trace-class operators. If $k(y, x)$ is a smooth function, then the operator $Tu(y) = \int k(y, x)u(x) dx$ is of trace class, and its trace is equal to $\int k(x, x) dx$, which can be regarded as the "sum" of the diagonal elements of the "continuous matrix" k .

2 Von Neumann Algebras

The *commutant* of a set S of bounded linear operators on a Hilbert space H is the collection S' of all operators on H that commute with every operator in the set S . The commutant of any set is an *algebra* of operators on H . That is, if T_1 and T_2 are in the commutant, then so are T_1T_2 and any linear combination $a_1T_1 + a_2T_2$.

As mentioned in the previous section, a *unitary representation* of a group G on a Hilbert space H is a collection of unitary operators U_g , labeled by elements of G , with the property that for any two group elements g_1 and g_2 the composition $U_{g_1}U_{g_2}$ is equal to $U_{g_1g_2}$. A *von Neumann algebra* is any algebra of operators on a complex Hilbert space H which is the commutant of some unitary representation of a group on H . Every von Neumann algebra is closed under adjoints and under limits of nearly every sort. For example, it is closed under pointwise limits: if $\{T_n\}$ is a sequence of operators in a von Neumann algebra M , and if $T_nv \rightarrow Tv$, for every vector $v \in H$, then $T \in M$.

It is easy to check that every von Neumann algebra M is equal to its own double commutant M'' (the commutant of the commutant of M). Von Neumann proved that if a self-adjoint algebra M of operators is closed under pointwise limits, then M is equal to the commutant of the group of unitary operators in its commutant, and is therefore a von Neumann algebra.

2.1 Decomposing Representations

Let $g \rightarrow U_g$ be a unitary representation of a group G on a Hilbert space H . If a closed subspace H_0 of H is mapped into itself by all the operators U_g , then it is said to be an *invariant subspace* for the representation. If H_0 is invariant, then since the operators U_g map H_0 to itself, their restrictions to H_0 constitute another representation of G , called a *subrepresentation* of the original.

PUP: 'into' is correct here.

A subspace H_0 is invariant for a representation, and so determines a subrepresentation, if and only if the orthogonal projection operator $P : H \rightarrow H_0$ belongs to the commutant of that representation. This points to a close connection between subrepresentations and von Neumann algebras. In fact, von Neumann algebra theory can be thought of as the study of the ways in which unitary representations can be decomposed into subrepresentations.

A representation is *irreducible* if it has no nontrivial invariant subspace. A representation that does have a nontrivial invariant subspace H_0 can be divided into two subrepresentations: those associated with H_0 and those associated with its orthogonal complement H_0^\perp . Unless both the representations H_0 and H_0^\perp are irreducible, we will be able to divide one or both of them into still smaller pieces by repeating the process that was just carried out for H . If the initial Hilbert space H is finite dimensional, then continuing in this way we will eventually decompose it into irreducible subrepresentations. In the language of matrices, we will obtain a basis for H with respect to which all the operators in the group are simultaneously block diagonal, in such a way that each block represents an irreducible group of unitary operators on a smaller Hilbert space.

Reducing a unitary representation on a finite-dimensional Hilbert space into irreducible subrepresentations is a bit like decomposing an integer into a product of prime factors. As with prime factorization, the decomposition process for a finite-dimensional unitary representation has only one possible end: there is, up to ordering, a unique list of irreducible representa-

tions into which a given unitary representation decomposes. But in infinite dimensions the decomposition process faces a number of difficulties, the most surprising of which is that there may be two decompositions of the same representation into entirely different sets of irreducible subrepresentations.

In the face of this, a different form of decomposition suggests itself, which is roughly analogous to the factorization of an integer into prime powers instead of individual primes. Let us refer to the prime powers into which an integer is decomposed as its *components*. They have two characteristic properties: no two components share a common factor, and any two (proper) factors of the same component *do* share a common factor. Similarly, one can decompose a unitary representation into *isotypical components*, which have analogous properties: no two distinct isotypical components share a common (meaning isomorphic) subrepresentation, and any two subrepresentations of the same isotypical component have themselves a common sub-subrepresentation. Any unitary representation (finite dimensional or not) can be decomposed into isotypical components, and this decomposition is unique.

In finite dimensions, every isotypical representation decomposes into a (finite) number of *identical* irreducible subrepresentations (like the prime factors of a prime power). In infinite dimensions this is not so. In effect, much of von Neumann algebra theory is concerned with analyzing the many possibilities that arise.

2.2 Factors

The commutant of an isotypical unitary representation is called a *factor*. Concretely, a factor is a von Neumann algebra M whose *center*, the set of all operators in M that commute with every member of M , consists of nothing more than scalar multiples of the identity operator. This is because projections in the center of M correspond to projections onto combinations of isotypical subrepresentations. Every von Neumann algebra can be uniquely decomposed into factors.

A factor is said to be of *type I* if it arises as the commutant of an isotypical representation that is a multiple of a single irreducible representation. Every type I factor is isomorphic to the algebra of all bounded operators on a Hilbert space. In finite dimensions, every factor is of type I, since as we already noted every isotypical representation decomposes into a multiple of one irreducible representation.

The existence of unitary representations with more than one decomposition into irreducible components is related to the existence of factors that are *not* of type I. Von Neumann, together with Francis Murray, investigated this possibility in a series of papers that mark the foundation of operator algebra theory. They introduced an order structure on the collection of subrepresentations of a given isotypical representation or, to put it in terms of the commutant, on the collection of projections in a given factor. If H_0 and H_1 are subrepresentations of the isotypical representation H , then we write $H_0 \leq H_1$ if H_0 is isomorphic to a subrepresentation of H_1 . Murray and von Neumann proved that this is a total ordering: either $H_0 \leq H_1$; or $H_1 \leq H_0$; or both, in which case H_0 and H_1 are isomorphic. For example, in a finite-dimensional type I situation, where H is a multiple of n copies of a single irreducible representation, each subrepresentation is the sum of $m \leq n$ copies of the irreducible representation, and the order structure of the (isomorphism classes of) subrepresentations is the same as the order structure of the integers $\{0, 1, \dots, n\}$.

Murray and von Neumann showed that the only order structures that can arise from factors are the following very simple ones:

Type I, $\{0, 1, 2, \dots, n\}$ or $\{0, 1, 2, \dots, \infty\}$;

Type II, $[0, 1]$ or $[0, \infty]$;

Type III, $\{0, \infty\}$.

The *type* of a factor is determined from the order structure of its projections according to this table.

In the case of factors of type II, the order structure is that of an interval of *real numbers*, not integers. Any subrepresentation of an isotypical representation of type II can be divided into yet smaller subrepresentations: we shall never reach an irreducible “atom.” Nevertheless, subrepresentations can still be compared in size by means of the “real-valued dimension” provided by Murray and von Neumann’s theorem.

A notable example of a factor of type II may be obtained as follows. Let G be a group and let $H = \ell^2(G)$ be a Hilbert space having basis vectors $[g]$ corresponding to the elements $g \in G$. Then there is a natural representation of G on H derived from the group multiplication law, called the *regular representation*: given an element g of G , the corresponding unitary map U_g is the linear operator that takes each basis vector $[g']$ in $\ell^2(G)$ to the basis vector $[gg']$. The commutant of this representation is a von Neumann algebra M . If G is a commutative group, then all the operators U_g are

in the center of M ; but if G is far enough from commutativity (for instance, if it is a free group), then M will have trivial center and will therefore be a factor. It can be shown that this factor is of type II. There is a simple explicit formula for the real-valued dimension of a subrepresentation corresponding to an orthogonal projection $P \in M$. Represent P by an infinite matrix relative to the basis $\{[g]\}$ of H . Because P commutes with the representation, it is easy to see that the diagonal elements of P are all the same, equal to some real number between 0 and 1. This real number is the dimension of the subrepresentation corresponding to P .

More recently, the Murray-von Neumann dimension theory has found unexpected applications in TOPOLOGY [I.3 §6.4]. Many important topological concepts, such as Betti numbers, are defined as the (integer-valued) dimensions of certain vector spaces. Using von Neumann algebras, one can define real-valued counterparts of these quantities that have useful additional properties. In this way, one can use von Neumann algebra theory to obtain topological conclusions. The von Neumann algebras used here are typically obtained by the construction of the previous paragraph from the FUNDAMENTAL GROUP [IV.6 §2] of some compact space.

2.3 Modular Theory

Type III factors remained rather mysterious for a long time; indeed, Murray and von Neumann were at first unable to determine whether any such factors existed. They eventually managed to do so, but the fundamental breakthrough in the area came well after their pioneering work, when it was realized that each von Neumann algebra has a special family of symmetries, its so-called *modular automorphism group*.

To explain the origins of modular theory, let us consider once again the von Neumann algebra obtained from the regular representation of a group G . We defined the operators U_g on $\ell^2(G)$ by multiplying *on the left* by elements of G ; but we could equally well have considered a representation defined by multiplying *on the right*. This would have yielded a different von Neumann algebra.

So long as we deal only with discrete groups G this difference is unimportant, because the map $S : [g] \mapsto [g^{-1}]$ is a unitary operator on H that interchanges the left and right regular representations. But for certain *continuous* groups the problem arises that the function $f(g)$ may be square-integrable while $f(g^{-1})$ is not. In this situation there is no simple unitary isomorphism

analogous to the one for discrete groups. To remedy this, one must introduce a correction factor called the *modular function* of G .

The project of modular theory is to show that something analogous to the modular function can be constructed for any von Neumann algebra. This object then serves as an invariant for all factors of type III, whether or not they are explicitly derived from groups.

Modular theory exploits a version of the GNS construction (section 1.4). Let M be a self-adjoint algebra of operators. A linear functional $\phi : M \rightarrow \mathbb{C}$ is called a *state* if it is positive in the sense that $\phi(T^*T) \geq 0$, for every $T \in M$ (this terminology is derived from the connection described earlier between Hilbert space theory and quantum mechanics). For the purposes of modular theory we restrict attention to *faithful* states, those for which $\phi(T^*T) = 0$ implies $T = 0$. If ϕ is a state, then the formula

$$\langle T_1, T_2 \rangle = \phi(T_1^* T_2)$$

defines an inner product on the vector space M . Applying the GNS procedure, we obtain a Hilbert space H_M . The first important fact about H_M is that every operator T in M determines an operator on H_M . Indeed, a vector $V \in H_M$ is a limit $V = \lim_{n \rightarrow \infty} V_n$ of elements in M , and we can apply an operator $T \in M$ to the vector V using the formula

$$TV = \lim_{n \rightarrow \infty} TV_n,$$

where on the right-hand side we use multiplication in the algebra M . Because of this observation, we can think of M as an algebra of operators on H_M , rather than as an algebra of operators on whatever Hilbert space we began with.

Next, the adjoint operation equips the Hilbert space H_M with a natural “antilinear” operator $S : H_M \rightarrow H_M$ by the formula¹ $S(V) = V^*$. Since $U_g^* = U_{g^{-1}}$ for the regular representation, this is indeed analogous to the operator S we encountered in our discussion of continuous groups. The important theorem of Minoru Tomita and Masamichi Takesaki asserts that, as long as the original state ϕ satisfies a continuity condition, the *complex powers* $U_t = (S^*S)^{it}$ have the property that $U_t M U_{-t} = M$, for all t .

The transformations of M given by the formula $T \mapsto U_t T U_{-t}$ are called the *modular automorphisms* of M . Alain Connes proved that they depend only in a rather inessential way on the original faithful state ϕ . To

PUP: Tim wants these words to stay. OK?

1. The interpretation of this formula on the completion H_M of M is a delicate matter.

be precise, changing ϕ changes the modular automorphisms only by *inner automorphisms*, that is, transformations of the form $T \mapsto UTU^{-1}$, where U is a unitary operator in M itself. The remarkable conclusion is that every von Neumann algebra M has a canonical one-parameter group of “outer automorphisms,” which is determined by M alone and not by the state ϕ that is used to define it.

The modular group of a type I or type II factor consists only of the identity transformation; however, the modular group of a type III factor is much more complex. For example, the set

$$\{t \in \mathbb{R} : T \mapsto U_t T U_t^{-1} \text{ is an inner automorphism}\}$$

is a subgroup of \mathbb{R} and an invariant of M that can be used to distinguish between uncountably many different type III factors.

2.4 Classification

A crowning achievement of von Neumann algebra theory is the classification of factors that are *approximately finite dimensional*. These are the factors that are in a certain sense limits of finite-dimensional algebras. Besides the range of the dimension function, which separates factors into types, the sole invariant is the *modular*. This is a flow on a certain space that is assembled from the modular automorphism group.

A lot of attention is currently being given to the long-standing problem of distinguishing among the type II factors associated with the regular representations of groups. Of special interest is the case of FREE GROUPS [IV.10 §2], around which has flourished the subject of free probability theory. Despite intensive effort, some fundamental questions remain open: at the time of writing it is unknown whether the factors associated with the free groups on two and on three generators are isomorphic.

Another important development has been *subfactor theory*, which attempts to classify the ways in which factors can be realized within other factors. A remarkable and surprising theorem of Vaughan Jones shows that, in the type II situation, where continuous values of dimensions are the norm, the dimensions of subfactors can in certain situations assume only a discrete range of values. The combinatorics associated with this result have also appeared in other apparently quite unrelated parts of mathematics, notably KNOT THEORY [III.46].

3 C*-Algebras

Von Neumann algebra theory helps describe the structure of a single representation of a group on a Hilbert space. But in many situations it is of interest to gain an understanding of all possible unitary representations. To shed some light on this problem we turn to a related but different part of operator algebra theory.

Consider the collection $\mathcal{B}(H)$ of all bounded operators on a Hilbert space H . It has two very different structures: *algebraic* operations, such as addition, multiplication, and formation of adjoints; and *analytic* structures, such as the operator norm

$$\|T\| = \sup\{\|Tu\| : \|u\| \leq 1\}.$$

These structures are not independent of one another. Suppose, for instance, that $\|T\| < 1$ (an analytic hypothesis). Then the geometric series

$$S = I + T + T^2 + T^3 + \cdots$$

converges in $\mathcal{B}(H)$, and its limit S satisfies

$$S(I - T) = (I - T)S = I.$$

It follows that $I - T$ is invertible in $\mathcal{B}(H)$ (an algebraic conclusion). One can easily deduce from this that the *spectral radius* $r(T)$ of any operator T (defined to be the greatest absolute value of any complex number in the spectrum of T) is less than or equal to its norm.

The remarkable *spectral radius formula* goes much further in the same direction. It asserts that $r(T) = \lim_{n \rightarrow \infty} \|T^n\|^{1/n}$. If T is *normal* ($TT^* = T^*T$), and in particular if T is self-adjoint, then it may be shown that $\|T^n\| = \|T\|^n$. As a result, the spectral radius of T is precisely equal to the norm of T . There is therefore a very close connection between the algebraic structure of $\mathcal{B}(H)$, particularly algebraic structure related to the adjoint operation, and the analytic structure.

Not all the properties of $\mathcal{B}(H)$ are relevant to this connection between algebra and analysis. A *C*-algebra* A is an abstract structure that has enough properties for the argument of the previous two paragraphs to remain valid. A detailed definition would be out of place here, but it is worth mentioning that a crucial condition relating norm, multiplication, and $*$ -operation is

$$\|a^*a\| = \|a\|^2, \quad a \in A,$$

called the *C*-identity* for A . We also note that special classes of operators on Hilbert space (unitaries, orthogonal projections, and so on) all have their counterparts in a general C*-algebra. For example, a *unitary* $u \in A$ satisfies $uu^* = u^*u = 1$, and a *projection* p satisfies $p = p^2 = p^*$.

A simple example of a C^* -algebra is obtained by starting with a single operator $T \in \mathcal{B}(H)$. The collection of all operators $S \in \mathcal{B}(H)$ that can be obtained as limits of polynomials in T and T^* is a C^* -algebra said to be *generated* by T . The C^* -algebra generated by T is commutative if and only if T is normal; this is one reason for the importance of normal operators.

3.1 Commutative C^* -Algebras

If X is a COMPACT [III.9] TOPOLOGICAL SPACE [III.92], then the collection $C(X)$ of continuous functions $f : X \rightarrow \mathbb{C}$ comes with natural algebraic operations (inherited from the usual ones on \mathbb{C}) and a norm $\|f\| = \sup\{|f(x)| : x \in X\}$. In fact, these operations make $C(X)$ into a C^* -algebra. The multiplication in $C(X)$ is *commutative*, because the multiplication of complex numbers is commutative.

A basic result of Gelfand and Naimark asserts that every commutative C^* -algebra is isomorphic to some $C(X)$. Given a commutative C^* -algebra A , one constructs X as the collection of all algebra homomorphisms $\xi : A \rightarrow \mathbb{C}$, and the *Gelfand transform* then associates with $a \in A$ the function $\xi \mapsto \xi(a)$ from X to \mathbb{C} .

The Gelfand–Naimark theorem is a foundational result of operator theory. For example, a modern proof of the spectral theorem might proceed as follows. Let T be a self-adjoint or normal operator on a Hilbert space H , and let A be the commutative C^* -algebra generated by T . By the Gelfand–Naimark theorem, A is isomorphic to $C(X)$ for some space X , which may in fact be identified with the spectrum of T . If v is a unit vector in H , then the formula $S \mapsto \langle Sv, v \rangle$ defines a state ϕ on A . The GNS space associated with this state is a Hilbert space of functions on X , and elements of $A = C(X)$ act as multiplication operators. In particular, T acts as a multiplication operator. A small additional argument shows that T is unitarily equivalent to this multiplication operator, or at least to a direct sum of such operators (which is itself a multiplication operator on a larger space).

Continuous functions can be composed: if f and g are continuous functions (with the range of g contained in the domain of f), then $f \circ g$ is also a continuous function. Since the Gelfand–Naimark theorem tells us that any self-adjoint element of a C^* -algebra A sits inside an algebra isomorphic to the continuous functions on the spectrum of a , we conclude that if $a \in A$ is self-adjoint, and if f is a continuous function defined on the

spectrum of a , then an operator $f(a)$ exists in A . This *functional calculus* is a key technical tool in C^* -algebra theory. For example, suppose that $u \in A$ is unitary and $\|u - 1\| < 2$. Then the spectrum of u is a subset of the unit circle in \mathbb{C} that does not contain -1 . One can define a continuous branch of the complex logarithm function on such a subset, and it follows that there is an element $a = \log u$ of the algebra such that $a = -a^*$ and $u = e^a$. The path $t \mapsto e^{ta}$, $0 \leq t \leq 1$, is then a continuous path of unitaries in A connecting u to the identity. Thus every unitary sufficiently close to the identity is connected to the identity by a unitary path.

3.2 Further Examples of C^* -Algebras

3.2.1 The Compact Operators

An operator on a Hilbert space has *finite rank* if its range is a finite-dimensional subspace. The operators of finite rank form an algebra, and its closure is a C^* -algebra called the algebra of *compact* operators and denoted \mathcal{K} . One can also view \mathcal{K} as a “limit” of matrix algebras

$$M_1(\mathbb{C}) \rightarrow M_2(\mathbb{C}) \rightarrow M_3(\mathbb{C}) \rightarrow \cdots,$$

where each matrix algebra is included in the next by

$$A \mapsto \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}.$$

Many natural operators are compact, including the integral operators that arose in Fredholm’s theory. The identity operator on a Hilbert space is compact if and only if that Hilbert space is finite dimensional.

3.2.2 The CAR Algebra

The presentation of \mathcal{K} as a limit of matrix algebras leads one to consider other “limits” of a similar sort. (We shall not attempt a formal definition of these limits here, but it is important to note that the limit of a sequence $A_1 \rightarrow A_2 \rightarrow A_3 \rightarrow \cdots$ depends on the homomorphisms $A_i \rightarrow A_{i+1}$ as well as on the algebras A_i .) One particularly important example is obtained as the limit

$$M_1(\mathbb{C}) \rightarrow M_2(\mathbb{C}) \rightarrow M_4(\mathbb{C}) \rightarrow \cdots,$$

where each matrix algebra is included in the next by

$$A \mapsto \begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix}.$$

This is called the *CAR algebra*, because it contains elements that represent the *canonical anticommutation relations* that arise in quantum theory. C^* -algebras find

several applications to quantum field theory and quantum statistical mechanics which extend von Neumann's formulation of quantum theory in terms of Hilbert space.

3.2.3 Group C^* -Algebras

If G is a group and $g \mapsto U_g$ is a unitary representation of G on a Hilbert space H , we can consider the smallest C^* -algebra of operators on H containing all the U_g ; this is called the C^* -algebra *generated* by the representation. An important example is the *regular representation* on the Hilbert space $\ell^2(G)$ generated by G , which we defined in section 2.2. The C^* -algebra that it generates is denoted $C_r^*(G)$. The subscript “r” refers to the regular representation. Considering other representations leads to other, potentially different, group C^* -algebras.

Consider, for example, the case $G = \mathbb{Z}$. Since this is a commutative group, its C^* -algebra is also commutative, and thus it is isomorphic to $C(X)$ for a suitable X , by the Gelfand–Naimark theorem. In fact, X is the unit circle S^1 , and the isomorphism

$$C(S^1) \cong C_r^*(\mathbb{Z})$$

takes a function on the circle to its Fourier series.

States defined on group C^* -algebras correspond to positive-definite functions defined on groups, and hence to unitary group representations. In this way new representations may be constructed and studied. For example, using states of group C^* -algebras it is possible to give to the set of irreducible representations of G the structure of a topological space.

3.2.4 The Irrational Rotation Algebra

The algebra $C^*(\mathbb{Z})$ is generated by a single unitary element U (corresponding to $1 \in \mathbb{Z}$). Moreover, it is the *universal example* of such a C^* -algebra, which is to say that given any C^* -algebra A and unitary $u \in A$, there is one and only one homomorphism $C^*(\mathbb{Z}) \rightarrow A$ sending U to u . In fact, this is nothing other than the functional calculus homomorphism for the unitary u .

If instead we consider the universal example of a C^* -algebra generated by *two* unitaries U, V subject to the relation

$$UV = e^{2\pi i\alpha} VU,$$

where α is irrational, we obtain a noncommutative C^* -algebra called the *irrational rotation algebra* A_α . The irrational rotation algebras have been studied intensively from a number of points of view. Using K -theory

(see below) it has been shown that A_{α_1} is isomorphic to A_{α_2} if and only if $\alpha_1 \pm \alpha_2$ is an integer.

It can be shown that the irrational rotation algebra is *simple*, which implies that *any* pair of unitaries U, V satisfying the commutation relation above will generate a copy of A_α . (Note the contrast with the case of a single unitary: 1 is a unitary operator, but it does not generate a copy of $C^*(\mathbb{Z})$.) This allows us to give a concrete representation of A_α on the Hilbert space $L^2(S^1)$, where U is the rotation through $2\pi\alpha$ and V is multiplication by $z : S^1 \rightarrow \mathbb{C}$.

4 Fredholm Operators

A *Fredholm operator* between Hilbert spaces is a bounded operator T for which the kernel and cokernel are finite dimensional. This means that the homogeneous equation $Tu = 0$ admits only finitely many linearly independent solutions, while the inhomogeneous equation $Tu = v$ admits a solution if v satisfies a finite number of linear conditions. The terminology arises from Fredholm's original work on integral equations; he showed that if K is an integral operator, then $I + K$ is a Fredholm operator.

For the operators that Fredholm considered, the dimensions of the kernel and cokernel must be equal, but in general this need not be so. The *unilateral shift operator* S , which maps the infinite “row vector” (a_1, a_2, a_3, \dots) to $(0, a_1, a_2, \dots)$, is an example. The equation $Su = 0$ has only the zero solution, but the equation $Su = v$ has a solution only if the first coordinate of the vector v is zero.

The *index* of a Fredholm operator is defined to be the integer difference

$$\text{index}(T) = \dim(\ker(T)) - \dim(\text{coker}(T)).$$

For example, every invertible operator is a Fredholm operator of index 0, whereas the unilateral shift is a Fredholm operator of index -1 .

4.1 Atkinson's Theorem

Consider the two systems of linear equations

$$\begin{cases} 2 \cdot 1x + y = 0 \\ 4x + 2y = 0 \end{cases} \quad \text{and} \quad \begin{cases} 2x + y = 0 \\ 4x + 2y = 0 \end{cases}.$$

Although the coefficients of these equations are very close, the dimensions of their kernels are quite different: the left-hand system has only the zero solution, whereas the right-hand system has the nontrivial solutions $(t, -2t)$. Thus the dimension of the kernel is an

unstable invariant of the system of equations. A similar remark applies to the dimension of the cokernel. By contrast, the index is stable, despite its definition as the difference of two unstable quantities.

An important theorem of Frederick Atkinson gives precise expression to these stability properties. Atkinson's theorem asserts that an operator T is Fredholm if and only if it is invertible modulo compact operators. This implies that any operator that is sufficiently close to a Fredholm operator is itself a Fredholm operator with the same index, and that if T is a Fredholm operator and K is a compact operator, then $T + K$ is a Fredholm operator with the same index as T . Notice that, since integral operators are compact operators, this contains Fredholm's original theorem as a special case.

4.2 The Toeplitz Index Theorem

TOPOLOGY [I.3 §6.4] studies those properties of mathematical systems that remain the same when the system is (continuously) perturbed. Atkinson's theorem tells us that the Fredholm index is a topological quantity. In many contexts it is possible to obtain a formula for the index of a Fredholm operator in terms of other, apparently quite different, topological quantities. Formulas of this sort often indicate deep connections between analysis and topology and often have powerful applications.

The simplest example involves the *Toeplitz operators*. A Toeplitz operator has a matrix with the special form

$$T = \begin{pmatrix} b_0 & b_1 & b_2 & b_3 & \cdots \\ b_{-1} & b_0 & b_1 & b_2 & \cdots \\ b_{-2} & b_{-1} & b_0 & b_1 & \cdots \\ b_{-3} & b_{-2} & b_{-1} & b_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

In other words, as you go down each diagonal of the matrix, the entries remain constant. The sequence of coefficients $\{b_n\}_{n=-\infty}^{\infty}$ defines a function $f(z) = \sum_{n=-\infty}^{\infty} b_n z^{-n}$ on the unit circle in the complex plane, called the *symbol* of the Toeplitz operator. It can be shown that a Toeplitz operator whose symbol is a continuous function which is never zero is Fredholm. What is its index?

The answer is given by thinking about the symbol as a mapping from the unit circle to the nonzero complex numbers: in other words, as a closed path in the nonzero complex plane. The fundamental topological

invariant of such a path is its *winding number*: the number of times it "goes around" the origin in the counter-clockwise direction. It can be proved that the index of a Toeplitz operator with nonzero symbol f is minus the winding number of f . For example, if f is the function $f(z) = z$ (with winding number $+1$), then the associated Toeplitz operator is the unilateral shift S that we encountered earlier (with index -1). The Toeplitz index theorem is a very special case of THE ATIYAH-SINGER INDEX THEOREM [V.2], which gives a topological formula for the indices of various Fredholm operators that arise in geometry.

4.3 Essentially Normal Operators

Atkinson's theorem suggests that compact perturbations of an operator are in some sense "small." This leads to the study of properties of an operator that are preserved by compact perturbation. For instance, the *essential spectrum* of an operator T is the set of complex numbers λ for which $T - \lambda I$ fails to be Fredholm (that is, invertible modulo compact operators). Two operators T_1 and T_2 are *essentially equivalent* if there is a unitary operator U such that UT_1U^* and T_2 differ by a compact operator. A beautiful theorem originally due to WEYL [VI.80] asserts that two self-adjoint or normal operators are essentially equivalent if and only if they have the same essential spectrum.

One might argue that the restriction to normal operators in this theorem is inappropriate. Since we are concerned with properties that are preserved by compact perturbation, would it not be more appropriate to consider *essentially normal* operators—that is, operators T for which $T^*T - T^*T$ is compact? This apparently modest variation leads to an unexpected result. The unilateral shift S is an example of an essentially normal operator. Its essential spectrum is the unit circle, as is the essential spectrum of its adjoint; however, S and S^* cannot be essentially equivalent, because S has index -1 and S^* has index $+1$. Thus some new ingredient, beyond the essential spectrum, is needed to classify essentially normal operators. In fact, it follows easily from Atkinson's theorem that if essentially normal operators T_1 and T_2 are to be essentially equivalent, then not only must they have the same essential spectrum but also, for every λ not in the essential spectrum, the Fredholm index of $T_1 - \lambda I$ must be equal to the Fredholm index of $T_2 - \lambda I$. The converse of this statement was proved by Larry Brown, Ron Douglas, and Peter Fillmore in the 1970s, using entirely novel techniques

that led to a new era of interaction between C^* -algebra theory and topology.

4.4 K -Theory

A remarkable feature of the Brown-Douglas-Fillmore work was the appearance within it of tools from ALGEBRAIC TOPOLOGY [IV.6], notably K -theory. Remember that, according to the Gelfand-Naimark theorem, the study of (suitable) topological spaces and the study of commutative C^* -algebras are one and the same; all the techniques of topology can be transferred, via the Gelfand-Naimark isomorphism, to commutative C^* -algebras. Having made this observation, it is natural to ask which of these techniques can be extended further, to provide information about *all* C^* -algebras, commutative or not. The first and best example is K -theory.

In its most basic form, K -theory associates with each C^* -algebra A an Abelian group $K(A)$, and with each homomorphism of C^* -algebras a corresponding homomorphism of Abelian groups. The building blocks for $K(A)$ can be thought of as generalized Fredholm operators associated with A ; the generalization is that these operators act on “Hilbert spaces” in which the complex scalars are replaced by elements of the C^* -algebra A . The group $K(A)$ itself is defined to be the collection of connected components of the space of all such generalized Fredholm operators. Thus if $A = \mathbb{C}$, for instance (so that we are dealing with classical Fredholm operators), then $K(A) = \mathbb{Z}$. This follows from the fact that two Fredholm operators are connected by a path of Fredholm operators if and only if they have the same index.

One of the great strengths of K -theory is that K -theory classes may be constructed from a variety of different ingredients. For example, every projection $p \in A$ defines a class in $K(A)$ which can be thought of as a “dimension” for the range of p . This connects K -theory to the classification of factors (section 2.2), and has become an important tool in the effort to classify various families of C^* -algebras, such as the irrational rotation algebras. (It was at one time thought that the irrational rotation algebras might not contain any nontrivial projections at all: the construction of such projections by Marc Rieffel was an important step in the development of C^* -algebra K -theory.) Another beautiful example is George Elliott’s classification theorem for locally finite-dimensional C^* -algebras like the CAR algebra; they are completely determined by K -theoretic invariants.

The problem of computing the K -theory groups of noncommutative C^* -algebras, particularly group C^* -algebras, has turned out to have important connections with topology. In fact, some key advances in topology have come from C^* -algebra theory in this way, thereby allowing operator algebraists to repay some of the debt they owe to the topologists for K -theory. The principal organizing problem in this area is the *Baum-Connes conjecture*, which proposes a description of the K -theory of group C^* -algebras in terms of invariants familiar in algebraic topology. Most of the progress on the conjecture to date is the result of work of Gennadi Kasparov, who dramatically broadened the original discoveries of Brown, Douglas, and Fillmore to cover not just single essentially normal operators but also noncommuting systems of operators, that is, C^* -algebras. Kasparov’s work is now a central component of operator algebra theory.

5 Noncommutative Geometry

DESCARTES’S [VI.11] invention of coordinates showed that one can do geometry by thinking about coordinate functions rather than directly thinking about points in space and their interrelationships: these coordinate functions are the familiar x , y , and z . The Gelfand-Naimark theorem can be viewed as one expression of this idea of passing from the “point picture” of a space X to the “field picture” of the algebra $C(X)$ of functions on it. The success of K -theory in operator algebras invites us to ponder whether the field picture might be *more powerful* than the point picture, since K -theory can be applied to noncommutative C^* -algebras which may not have any “points” (homomorphisms to \mathbb{C}) at all.

One of the most exciting research frontiers in operator algebra theory is reached along a path which develops these thoughts. The *noncommutative geometry* program of Connes takes seriously the idea that a general C^* -algebra should be thought of as an algebra of functions on a “noncommutative space,” and goes on to develop “noncommutative” versions of many ideas from geometry and topology, as well as completely new constructions that have no commutative counterpart. Noncommutative geometry begins with the creative reformulation of ideas from ordinary geometry in ways that involve only operators and functions, but not points.

Consider, for instance, the circle S^1 . The algebra $C(S^1)$ reflects all the topological properties of S^1 , but to

incorporate its *metric* (distance-related) properties as well we look not just at $C(S^1)$ but at the pair consisting of the algebra $C(S^1)$ and the operator $D = i d/d\theta$ on the Hilbert space $H = L^2(S^1)$. Notice that if f is a function on the circle (considered as a multiplication operator on H), then the commutator $Df - fD$ is also a multiplication operator, this time by $idf/d\theta$. It follows that ordinary measurements of angular distance between points on the circle can be recovered from $C(S^1)$ and D by the formula

$$d(p, q) = \max\{|f(p) - f(q)| : \|Df - fD\| \leq 1\}.$$

Connes argues that operator $|D|^{-1}$ plays the role of the “unit of arc-length ds ” in this and many other, more complicated situations.²

Another feature of the examples Connes considers, also of central importance in noncommutative geometry, is the fact that the operator $|D|^{-k}$ is a trace-class operator (see section 1.5) when k is large enough. In the case of the circle, k needs to be bigger than 1. Computations with traces connect noncommutative geometry to COHOMOLOGY THEORY [IV.6 §4]. We now have two kinds of “noncommutative algebraic topology,” namely K -theory and a new variant of homology called *cyclic cohomology*; the connection between the two is provided by a very general index theorem.

There are several procedures that produce noncommutative C^* -algebras (to which Connes’s methods can be applied) from classical geometric data. The irrational rotation algebras A_θ are examples; the classical picture to which they apply is the QUOTIENT SPACE [I.3 §3.3] of the circle by the group of rotations through multiples of θ . Classical methods of geometry and topology are unable to handle this quotient space, but the noncommutative approach via A_θ is much more successful.

An exciting but speculative possibility is that the basic laws of physics should be addressed from the perspective of noncommutative geometry. The transition to noncommutative C^* -algebras can be viewed as analogous to the transition from classical to quantum mechanics. However, Connes has argued that noncommutative C^* -algebras play a role in describing the physical world even before the transition is made to quantum physics.

2. The operator D is not quite invertible since it vanishes on constant functions. A small modification must therefore be made before considering inverse operators. The operator $|D|$ is by definition the positive square root of D^2 .

Further Reading

- Connes, A. 1995. *Noncommutative Geometry*. Boston, MA: Academic Press.
 Davidson, K. 1996. *C^* -Algebras by Example*. Providence, RI: American Mathematical Society.
 Fillmore, P. 1996. *A User’s Guide to Operator Algebras*. Canadian Mathematical Society Series of Monographs and Advanced Texts. New York: John Wiley.
 Halmos, P. R. 1963. What does the spectral theorem say? *American Mathematical Monthly* 70:241–47.

IV.16 Mirror Symmetry

Eric Zaslow

1 What Is Mirror Symmetry?

Mirror symmetry is a phenomenon found in theoretical physics that has had profound mathematical applications. It burst onto the mathematical scene after Candelas, de la Ossa, Green, and Parkes exploited the physical phenomenon to make precise predictions about certain sequences of numbers describing geometric spaces. The sequence predicted by those authors began 2875, 609 250, 317 206 375, ..., and was far beyond the scope of calculation at the time. The phenomenon of mirror symmetry is that some physical theories have equivalent, “mirror” theories that lead to the same predictions. If some prediction requires a hard calculation but is easy to perform in the mirror theory, then you can get the answer for free! These physical theories do not have to be realistic models of physics. For instance, beginning students of physics often study point particles on frictionless planes. Although they are unrealistic, such toy models can bring the physical concepts into focus and their analysis can give rise to very interesting mathematics.

1.1 Exploiting Equivalences

Children at school in the 1950s used slide rules to exploit the equivalence of multiplication of positive numbers with addition of real numbers. Given the problem of multiplying two large numbers a and b , they would use a table to look up the logarithms $\log(a)$ and $\log(b)$ (to a certain number of significant figures), then add them by hand. They would then use the same table to find which number had a logarithm equal to $\log(a) + \log(b)$. The answer is ab .

College students sometimes exploit the equivalence defined by FOURIER TRANSFORMS [III.27] to solve differential equations. Basically, the Fourier transform is

a rule that maps one function $f(x)$ to a new function $\hat{f}(p)$. What is nice is that the transform of the derivative $f'(x)$ relates in a very simple way to $\hat{f}(p)$: it is $ip\hat{f}(p)$, where i is the imaginary number $\sqrt{-1}$. If you want to solve a differential equation such as $f'(x) + 2f(x) = h(x)$, where $h(x)$ is a given function and you are trying to find f , you can map the equation to its Fourier transform equation $ip\hat{f}(p) + 2\hat{f}(p) = \hat{h}(p)$. This is much easier: it is an algebraic equation rather than a differential equation, and has the solution $\hat{f}(p) = \hat{h}(p)/(2 + ip)$. The solution $f(x)$ is then the function which has $\hat{h}(p)/(2 + ip)$ as its Fourier transform.

Mirror symmetry is like a fancy Fourier transform, mapping much more information than is contained in a single function. Every aspect of a physical theory is involved.

This article will (eventually) focus on the mathematics of mirror symmetry, but it is crucial to understand its physical origins. We therefore develop a guide to physics (see VERTEX OPERATOR ALGEBRAS [IV.17 §2] for further discussion of these topics). This is in no way an adequate treatment—a separate *Companion to Physics* would be needed—but we hope to give enough of the flavor of the subject to help the reader with the later sections. (A reader familiar with physical theories may wish to skip the next section and refer back as needed.)

2 Theories of Physics

2.1 Formulations of Mechanics and Action Principles

2.1.1 Newtonian Physics

Newton's second law states that a particle moving through space accelerates¹ in proportion to the force it experiences: $F = m\ddot{x}$. The force is itself the (negative) gradient of a gravitational potential $V(x)$, so this equation can be written $m\ddot{x} + \nabla V(x) = 0$. Stationary particles sit at minima of the potential: examples are a ball in equilibrium at the end of a spring, or a pea at the bottom of a bowl. In stable situations, there is a restoring force proportional to some displacement distance. This means that in some appropriate coordinate, $F \sim -x$, so $V(x) = kx^2/2$, for some k . The solutions are oscillatory, with period $\omega = \sqrt{k/m}$. This model is called the *simple harmonic oscillator*.

1. Acceleration is the second derivative of position with respect to time. We denote position by x , which is shorthand for a three-component position vector, and we denote time derivatives by dots, so acceleration is denoted by \ddot{x} .

2.1.2 The Least Action Principle

Every major theory can also be formulated by means of an idea known as the *least action principle*. Let us see how it works for the equations of Newtonian mechanics. Consider an arbitrary path of a particle $x(t)$ and form the quantity

$$S(x) = \int [\frac{1}{2}m\dot{x}^2 - V(x)] dt.$$

Here and below, the notation x may represent more than one coordinate. If x is used as a point in space-time, it will include the time coordinate, if that is not otherwise noted. Likewise, we omit component notation on most vectors. The notation should be clear from the context. The quantity $S(x)$, which is known as the *action*, equals the kinetic energy minus the potential energy. One then considers which paths minimize this action. That is, we ask which paths $x(t)$ have the property that, when they are perturbed by a small amount $\delta x(t)$, the action is unchanged, to leading order. (So in fact we require only that the action is unchanged to first order, and not that it is actually minimized. Solutions of saddle-point type are allowed.) The answer turns out to be precisely those paths that satisfy $m\ddot{x} + \nabla V(x) = 0$.²

For example, consider the simple harmonic oscillator in two dimensions. We can model x as a complex number and set $V(x) = k|x|^2$. The action is then $\int \frac{1}{2}[m|\dot{x}|^2 - k|x|^2]$. Note that a phase rotation $x \rightarrow e^{i\theta}x$ leaves the action invariant, and is therefore a symmetry of the equations of motion.

Lesson. Physical solutions extremize the action.

The principle of least action applies to many other physical situations, as we shall see below. First, though, we describe another formulation of mechanics.

2.1.3 The Hamiltonian Formulation of Mechanics

HAMILTON's [VI.37] formulation of the equations of motion also deserves mention. It leads to first-order equations. Let S be the action and define L by $S = \int L dt$, and consider the (typical) case where L is a function of coordinates x and their time derivatives \dot{x} . Then set $p = dL/d\dot{x}$, a function that can depend both on x and on \dot{x} . (In the example $L = \frac{1}{2}m\dot{x}^2 - V(x)$ that we have already considered, we find that $p = m\dot{x}$, or $\dot{x} = p/m$.)

2. To see this, replace x by $x + \delta x$ in the action and keep only the linear terms in δx and its time derivative. For V the linear terms are $(\nabla V)\delta x$. One then has to integrate by parts to remove the time derivative of δx and isolate it as a factor in the integrand. The integral will be zero for arbitrary variations δx only when the term multiplying it vanishes. This gives the equation. Try it!

Now let us consider the function $H = p\dot{x} - L$, which is called the HAMILTONIAN [III.35], and change variables from (x, \dot{x}) to (x, p) so as to remove all mention of \dot{x} . In the example, H works out to be

$$\frac{p^2}{m} - \left(\frac{p^2}{2m} - V(x) \right) = \frac{p^2}{2m} + V(x),$$

which is the total energy. For the simple harmonic oscillator, $H = p^2/2m + kx^2/2$.

The equations $\dot{x} = \partial H / \partial p$ and $\dot{p} = -\partial H / \partial x$ are the equations of motion in the Hamiltonian formulation; they can be shown to be equivalent to those obtained from the action principle. In the example, $\dot{x} = p/m$ and $\dot{p} = -\nabla V$. Using the first equation to replace p by $m\dot{x}$ in the second, we recover the equation $m\ddot{x} + \nabla V(x) = 0$. More generally, one can consider the time derivative of some quantity $f(x, p)$ constructed from p and x and prove—using the chain rule and the equations of motion—that

$$\dot{f} = \frac{\partial f}{\partial x} \frac{\partial H}{\partial p} - \frac{\partial f}{\partial p} \frac{\partial H}{\partial x} = \{H, f\}.$$

The term in the middle is called the *Poisson bracket* of H and f , denoted $\{H, f\}$.

Lesson. The Hamiltonian controls time dependence through the Poisson bracket.

Notice that when we plug the coordinates x and p themselves into the bracket, we derive the identity

$$\{x, p\} = -1. \quad (1)$$

It is also possible to begin with the Hamiltonian viewpoint. One considers a space endowed with a bracket operation on functions, such that there are coordinate functions (not uniquely determined) obeying $\{x, p\} = -1$. The mechanical model is defined by a function $H(x, p)$, which determines the dynamics.

2.1.4 Symmetry

A brief remark on symmetry is in order. NOETHER [VI.76] proved that in the action formulation of mechanics, a symmetry of the action results in a conserved quantity. The prototypical example is translational or rotational symmetry, where the potential of a particle is invariant under some direction of translation or rotation: the corresponding conserved quantity is then momentum or angular momentum. In the example above, $V(x) = k|x|^2/2$ is independent of θ , the phase of x . The equation of motion determined by varying θ is $d(m|x|^2\dot{\theta})/dt = 0$, so in this case it is the angular momentum $m|x|^2\dot{\theta}$ that is conserved. In

the Hamiltonian formulation, since a conserved quantity $f(x, p)$ does not change with time, it must have zero Poisson bracket with the Hamiltonian: $\{H, f\} = 0$. In particular, the Hamiltonian itself is conserved.

2.1.5 Action Functions for Other Theories

Returning now to action principles, we shall see how different physical theories are described through different actions. In electricity and magnetism, MAXWELL'S EQUATIONS [IV.13 §1.1] can be formulated in the form $\delta S = 0$, where now the action S takes the form of an integral over space and time of the electric (E) and magnetic (B) fields. In the case where there are no sources, the action is written

$$S = \frac{1}{8\pi e^2} \int [E^2 - B^2] dx dt, \quad (2)$$

where e is the electric charge of an electron. There is one important difference from the previous example, which is that the variations of the action must be taken with respect to the fundamental fields, and E and B are not fundamental as they are derived from the electromagnetic potential $A = (\phi, A)$ by the equations $E = \nabla\phi - \dot{A}$, $B = \nabla \times A$. If you rewrite S in terms of A , vary A by δA , and set $\delta S = 0$, then you recover Maxwell's equations from the least action principle.

It is clear that the electromagnetic action merely changes sign under the replacement $E \rightarrow B$, $B \rightarrow -E$, and therefore any solution $\delta S = 0$ remains a solution under the transformation. This is an example of an equivalence of a classical theory of physics. In fact, this symmetry extends to the case where there are sources (such as electrons) if we also interchange electric and magnetic sources. (No magnetic sources have been observed in the universe, but a theory with such objects still makes sense.)

Lesson. Physical equivalences act on fields and their sources.

Electricity and magnetism is a “field theory,” which means that the degrees of freedom involve functions that depend on position in space. Contrast this with Newtonian mechanics, where the spatial degrees of freedom are just the coordinates of the particle(s). However, there is not much conceptual distance between the two, as can be seen in the following toy model.

We will consider the simplest example: a scalar field, ϕ . That is, ϕ is just a function that takes numerical values. Now imagine that space has just one dimension, not three, and further that that dimension is a

circle, which we can describe with an angular coordinate, θ . At any fixed point in time we can use **FOURIER SERIES** [III.27] to write the scalar field as $\phi(\theta) = \sum_n c_n \exp(in\theta)$, where the c_n are the Fourier coefficients, and if we want the values of ϕ to be real numbers then we must insist that $c_{-n} = c_n^*$. We can then think of $\phi(\theta)$ not as a function but as an infinite-dimensional vector (c_0, c_1, \dots) . The spatial dependence of ϕ is completely determined by the coefficients c_n . If we now wish to consider time dependence, then all we have to do is use time-dependent components $(c_0(t), c_1(t), \dots)$, which looks a lot like an infinite set of quantum-mechanical particles c_n . Thus, the function ϕ has the Fourier expansion $\phi(\theta, t) = \sum_n c_n(t) \exp(in\theta)$.

The simplest action for a scalar field ϕ that allows wave-like solutions of the equations of motion serves as a natural analogue of equation (2):

$$S = \int \frac{1}{2\pi} [(\dot{\phi})^2 - (\phi')^2] d\theta dt, \quad (3)$$

where $\phi' = \partial\phi/\partial\theta$. When we plug the Fourier expansion into the action and perform the θ integration, we get

$$S = \int \sum_n [\dot{c}_n^2 - n^2 |c_n|^2] dt. \quad (4)$$

Note that the term in brackets is just the action for a particle c_n in a quadratic potential, as in section 2.1.2. We simply have an infinite number of harmonic oscillators (with the exception of the c_0 degree of freedom, which corresponds to a free particle in no potential).

Lesson. Field theory is like point particle theory with an infinite number of particles. The particles correspond to the degrees of freedom of the field. When the action is just quadratic in the derivatives, the particles have an interpretation as simple harmonic oscillators.

Even **GENERAL RELATIVITY** [IV.13] fits into this framework as a field theory. For a space-time M , the field is the **RIEMANNIAN METRIC** [I.3 §6.10] on space-time. The metric is what determines the lengths of paths between points—so a stretching of space-time, for example, is represented by a rescaled metric. The action is then constructed as the integral of the Riemannian curvature scalar R over space-time: $S = \int_M R$.³

2.2 Quantum Theory

Mirror symmetry is an equivalence of quantum theories, so we must develop an understanding of what a

quantum theory is and what an equivalence looks like. There are two formulations of quantum mechanics: the operator formulation and Feynman's path-integral formulation.

Both formulations are probabilistic, meaning that you cannot predict exactly what will be observed in a single measurement, but you can make precise predictions about what will be observed after multiple, repeated measurements in the same environment. For instance, your experimental apparatus may involve a beam of electrons hitting a screen and making a mark. The beam will contain millions of electrons, so the pattern of marks on the screen can be predicted with great accuracy. However, we cannot say what will happen to a single, given electron—all we can do is assign probabilities to the outcomes of various measurements. These probabilities are encoded in the so-called “wave function” Ψ of the particle.

2.2.1 Hamiltonian Formulation

In the operator formulation of quantum mechanics, the positions and momenta of classical mechanics (and any quantity formed from them) are converted into **OPERATORS** [III.52] acting on a **HILBERT SPACE** [III.37] according to the following rule: *replace the Poisson bracket $\{\cdot, \cdot\}$ by $i/\hbar[\cdot, \cdot]$* , where $[A, B] = AB - BA$ is the *commutator bracket* and \hbar is Planck's constant. Thus, for example, we get from equation (1) the relation $[x, p] = i\hbar$. The state of a particle (or system) is now defined not as a set of values of x and p but as a vector Ψ in the Hilbert space. Once again, time evolution is determined by the Hamiltonian, H , but now H is an operator. The basic dynamical equation is

$$H\Psi = i\hbar \frac{d}{dt} \Psi. \quad (5)$$

This is called the *Schrödinger equation*.

Lesson. To quantize a classical theory, replace ordinary degrees of freedom by operators on a vector space; replace Poisson brackets by commutator brackets.

In the case where we have a particle on the real line \mathbb{R} , the Hilbert space is the space of square-integrable functions $L^2(\mathbb{R})$, so we write Ψ as $\Psi(x)$. The commutation relation is obeyed if we think of x as the operator that sends the function $\Psi(x)$ to the function $x\Psi(x)$. Now the relation $[x, p] = i\hbar$ means that we should represent p as the operator $-i\hbar(d/dx)$. The values of the classical quantity associated with an operator correspond to the **EIGENVALUES** [I.3 §4.3] of that operator,

3. In 3-space, the paraboloid $z = \frac{1}{2}ax^2 + \frac{1}{2}by^2$ has curvature ab at the origin.

so for example a state with momentum p has the form $\Psi \sim \exp(ipx/\hbar)$. Unfortunately, this is not square-integrable on the real line, but it would become so if we identified x and $x + 2\pi R$, for some number (radius) $R > 0$. Topologically, this COMPACTIFIES [III.9] \mathbb{R} to a circle, but note that Ψ will be single-valued only if $p = n\hbar/R$, where n is an integer. Thus, momentum is “quantized” in units of \hbar/R .⁴ The integer label of the c_n of equation (4) can therefore also be thought of as a momentum.

In the above example, \mathbb{R} is the degree of freedom of the classical coordinate x . In other examples, there is a copy of $L^2(\mathbb{R})$ for each real degree of freedom, whether or not it represents a geometric location.

Another novelty is that position and momentum do not commute as operators in quantum mechanics, meaning they cannot be simultaneously diagonalized: you cannot specify the position and momentum simultaneously. This is a form of Heisenberg’s uncertainty principle (see OPERATOR ALGEBRAS [IV.15 §1.3]).

2.2.2 Symmetry

As the rules of quantization would suggest, a symmetry of a quantum theory is an operator A such that $[H, A] = 0$. That is, A commutes with the Hamiltonian, and therefore respects the dynamics.

2.2.3 Example: The Simple Harmonic Oscillator

We now discuss an example that will be useful later on for understanding quantum field theory and mirror symmetry: the simple harmonic oscillator in quantum mechanics. Suppose that the constants are chosen so that the Hamiltonian is given by $H = x^2 + p^2$. If one defines $a = (x + ip)/\sqrt{2}$ and $a^\dagger = (x - ip)/\sqrt{2}$, then one can show that a^\dagger raises the energy of a state by one unit⁵ and a lowers the energy by one unit. Invoking the physical argument that there is a ground state Ψ_0 of lowest energy, this state must obey $a\Psi_0 = 0$. One then finds that all states can be written in terms of the basis

4. We shall occasionally choose our units to make \hbar equal to 1. For example, we could work in the fictitious time unit of “sqeconds,” one second equals \hbar sqeconds.

5. Here is the calculation: $[a, a^\dagger] = 1$ and $H = a^\dagger a + \frac{1}{2}$. Further, $[H, a^\dagger] = a^\dagger$ and $[H, a] = -a$. These equations have the following interpretation. Suppose Ψ is an eigenvector of H with eigenvalue (energy) E . Then $H\Psi = E\Psi$. Consider $a^\dagger\Psi$. One quickly finds that

$$\begin{aligned} H(a^\dagger\Psi) &= (Ha^\dagger - a^\dagger H + a^\dagger H)\Psi = ([H, a^\dagger] + a^\dagger H)\Psi \\ &= (a^\dagger + a^\dagger E)\Psi = (E + 1)(a^\dagger\Psi). \end{aligned}$$

We learn that $a^\dagger\Psi$ has eigenvalue $E + 1$, so a^\dagger has “raised” the energy by one unit.

vectors $\Psi_n = (a^\dagger)^n\Psi_0$ with energy $n + \frac{1}{2}$. Note that Ψ_0 has energy $\frac{1}{2}$.⁶ The basis $\{\Psi_n\}$ is called the *occupation number* basis, since the interpretation is that Ψ_n has n energy “quanta” above the ground state.

2.2.4 Path-Integral Formulation

Feynman’s path integral formulation of quantum mechanics builds on the idea of the least action principle. In this formulation, the probability of an experiment is calculated through an average over *all* paths of particles, and not just the ones which extremize the action. Each path $x(t)$ is weighted by the factor $\exp(iS(x)/\hbar)$, where $S(x)$ is the action of the path $x(t)$ and \hbar is Planck’s constant, which is very small compared with macroscopic action scales. This average can be an imaginary number, but the probability of the process is the square of its absolute value.

Note that $\exp(iS/\hbar) = \cos(S/\hbar) + i\sin(S/\hbar)$, so if S changes appreciably when we vary $x(t)$, then the real and imaginary parts will oscillate rapidly, since \hbar is small. Then, when we integrate over paths $x(t)$, the positive and negative oscillations will roughly cancel. As a result, the main contributions to the weighted sum over paths will come from those paths for which S does not vary when the path does: the classical paths! However, if the variations are sufficiently small compared with \hbar , then nonclassical paths can contribute appreciably. One typically separates the degrees of freedom into the classical trajectory piece and the quantum fluctuations near it. Then one can organize the path integral in a perturbation theory around the parameter \hbar .

We have not yet discussed the integrand of the path integral, and will not go into the details of this. The main point is that the theory makes a prediction about the likelihood of measuring a physical process. Each process determines a possible integrand. For example, from our discussion above we learn that the integrand for measuring the likelihood of a quantum-mechanical particle going from the point x_0 at time t_0 to the point x_1 at time t_1 gives nonzero weight—determined by the exponentiated action—to all paths that go from x_0 to x_1 as t goes from t_0 to t_1 , and zero weight to all other paths.

It is illustrative to consider a toy model of a path integral on a “space-time” that consists of just a single point. Then the possible “paths” of a scalar field, say, are simply the values that the field can take at the point,

6. It is instructive to write these equations in terms of the operators defined by x and p .

so they are real numbers. The action is then an ordinary function $S(x)$ on \mathbb{R} . For the purposes of this example, let us consider the case where $iS/\hbar = -x^2 + \lambda x^3$. The possible integrands are (sums of) powers of x , so the basic path integrals to perform are $\int x^k \exp(-\frac{1}{2}x^2 + \lambda x^3) dx$, which we denote by $\langle x^k \rangle$. The value at $\lambda = 0$ is easily calculated.⁷ For small λ we expand $e^{\lambda x^3}$ as $1 + \lambda x^3 + \lambda^2 x^6/2 + \dots$, and evaluate each term by the same methods as for $\lambda = 0$. This is how we construct a well-defined perturbation theory, even when the integral is not calculable.

As we see from this example, path integrals are easiest when the action is only quadratic in the variables, just as we found in the operator formulation of quantum mechanics. The mathematical reason for this is that Gaussian integrals (exponentials of squares) can be done explicitly, while integrals involving exponentials of cubics or higher are difficult or impossible. For quadratic actions, the path integral can be evaluated exactly, but when cubic or higher terms appear, the perturbation series is necessary.

2.2.5 Quantum Field Theory

The generalization to field theories follows our earlier pattern. We think of quantum field theories, then, as being like quantum mechanics with infinite numbers of particles. In fact, the quantum field theories in which the fields ϕ and their derivatives do not have more than quadratic terms in the action are easily understood in this way—we had a preview of this in equation (4). The Fourier components correspond to particles indexed by their momenta. Each one looks like a simple harmonic oscillator at some frequency, which will depend on the Fourier coefficient. The quantum Hilbert space is then a (tensor) product of lots of different “occupation number Hilbert spaces,” one for each Fourier component of each field. Since the occupation number basis is also an energy eigenbasis, these states have a simple time evolution under the Hamiltonian H . That is, if $H = E$ on some state $\Psi(t = 0)$, then that state evolves like

$$\Psi(t) = \exp(iEt/\hbar)\Psi(0).$$

However, if the action includes terms that are *cubic or higher*, then things get interesting: particles can decay!

7. Consider

$$\begin{aligned} \int \exp(-\tfrac{1}{2}x^2 + Jx) dx &= \int \exp(-\tfrac{1}{2}(x + J)^2) \exp(J^2/2) dx \\ &= \sqrt{2\pi} \exp(J^2/2). \end{aligned}$$

Now if we differentiate this answer with respect to J , and set $J = 0$, we get $\langle x \rangle$. Taking k derivatives gives $\langle x^k \rangle$, and the theory is solved.

This can be seen, for example, from the scalar field of equation (3) if we include a term ϕ^3 in the action, and therefore also the Hamiltonian. If we write this using Fourier components, we get terms involving three oscillators, such as $a_3^\dagger a_4^\dagger a_7$. To see this, recall that after we quantize the real field ϕ , the Fourier components c_n act as harmonic oscillators, and we have written a_n for the associated creation and annihilation operators. Since the Hamiltonian governs time evolution according to equation (5), this means that over time one particle (the 7 mode) can decay into two others (the 3 and the 4). Such decay processes occur in real life, and it is a great triumph of quantum field theory that it can predict such events with astounding accuracy.

In fact, because the space of paths of fields is infinite dimensional, the path integral in quantum field theory is not usually defined in a mathematically rigorous way. However, the perturbation series for producing predictions can be defined just as for quantum mechanics, and this is how physicists make their predictions in practice. This perturbation series is organized in terms of *Feynman diagrams* (which are discussed in VERTEX OPERATOR ALGEBRAS [IV.17]). These diagrams, and the rules for computing them, completely solve the perturbation problem.

As in the example of quantum mechanics, different integrands of the path integral correspond to different predictions. If Φ is some function of the fields of some quantum field theory, we write $\langle \Phi \rangle$ for the path integral with Φ as an integrand (as we did for $\langle x^k \rangle$ in the previous section). We call such a term a “correlation function.” If $\Phi = \phi_1(x_1) \cdots \phi_n(x_n)$, the answer will depend on the action of the theory, the fields ϕ_i , and the space-time points x_i .

One might wonder if a symmetry of a classical theory always remains a symmetry of the same theory after quantization. The answer is sometimes no. Such a case is known as an “anomaly.” Roughly speaking, this is because the measure of integration of the path integral is not preserved under the symmetry, but this is a somewhat heuristic explanation because the path integral has no rigorous definition in general.

Returning to our cubic example, if the interaction term ϕ^3 has a coefficient λ , so that it is $\lambda\phi^3$, then we organize the perturbation series as a power series in λ . In terms of paths, probabilities of decay processes can be evaluated by considering paths that split into two—like the letter Y—with each leg carrying the label of the appropriate particle.

2.2.6 String Theory

Feynman's perturbation theory has an important generalization in *string theory*. String theory considers particles not as points but as loops. Instead of paths of particles through space-time, we get paths of loops, which look like two-dimensional surfaces. String theory amplitudes are computed by summing over all surfaces. These sums are organized in a perturbation series in powers of the so-called *string coupling constant*, λ_g . The power of λ_g in the perturbation series depends on the number of holes in the surface.

The surfaces are called *worldsheets*. At each point of the worldsheet, its location in space-time is determined by coordinates X^i . These coordinates themselves depend on the location on the worldsheet. In effect, we get an *auxiliary* theory: a field theory of coordinates on the two-dimensional surface! In string theory, even this two-dimensional field theory must be considered as a quantum field theory. The fields of the two-dimensional theory are maps from the surface to actual space-time. However, from the point of view of the worldsheet, the worldsheet itself is a two-dimensional space-time and the maps are fields on *this* space-time with values in some other (target) space.

Mirror symmetry was discovered while studying these quantum field theories on two-dimensional surfaces. Subsequently, the same phenomenon was discovered in the case where the strings were not closed loops but filaments with endpoints. Both cases play an important role below.

3 Equivalence in Physics

Mirror symmetry is a particular type of equivalence of quantum field theories. As we have seen, quantum field theories are rules for producing probabilities of physical processes. In the path-integral formulation, probabilities are computed from correlation functions of fields. According to Feynman, these correlation functions can be thought of as being averages over all paths of fields. Each path is weighted by $\exp(iS/\hbar)$, where S is the action of the path and \hbar is Planck's constant. Let us denote the correlation function of some integrand Φ in theory A as $\langle \Phi \rangle_A$. Recall that Φ can depend on various fields ϕ_i and points of space-time x_i , and the correlation function will depend on all these and the action of theory A .

Equivalence, then, is a map from all possible fields ϕ_i in a theory A to corresponding fields $\tilde{\phi}_i$ in a theory

B such that

$$\langle \Phi \rangle_A = \langle \tilde{\Phi} \rangle_B.$$

(For the moment, we deliberately neglect to notate the dependence on the points x_i .) One special correlation function is $\langle 1 \rangle$, which we call the *partition function* and denote by Z . As the field 1 always gets mapped to 1, we derive the corollary that the partition functions must be equal: $Z_A = Z_B$.

Of course, this all has a description in the operator formulation of the quantum theory. Each state Ψ and each operator a in one theory must get mapped to a corresponding state $\tilde{\Psi}$ and operator \tilde{a} in the mirror theory, in such a way that corresponding operators map corresponding states to states which themselves correspond. Here one sees the sharp analogy with the slide rule and the operations of multiplication and addition of numbers.

Each theory is typically described through some mathematical model, so an equivalence implies a host of mathematical identities between quantities constructed from corresponding models.

More specifically, mirror symmetry refers to an equivalence of quantum field theories on a two-dimensional surface. The most typical example of mirror symmetry is the physical theory whose fields are maps φ from a two-dimensional RIEMANN SURFACE [III.81] Σ to some target space, M . Such a theory is called a *sigma model*. As we saw above, in string theory M plays the role of actual space-time, but for our purposes we can even consider the case where M is the real line \mathbb{R} , so that φ is an ordinary function. This case has already been studied in section 2.1.5. The action is given in equation (4). We can then write the partition function as

$$Z = \langle 1 \rangle = \int [\mathcal{D}\varphi] e^{iS(\varphi)/\hbar},$$

where $[\mathcal{D}\varphi]$ represents the measure of integration over all paths.⁸

One approach to evaluating the partition function Z is through a process known as *Wick rotation*. One first Euclideanizes the time coordinate by writing $\tau = it$ (this is the Wick rotation), which leads to an imaginary Euclidean action iS_E . One then tries to evaluate the path integral in this framework, hoping that the answer will be HOLOMORPHIC [I.3 §5.6]. If it is, then one can use

8. Warning: these expressions represent only the "bosonic" part of a theory with "supersymmetry," meaning, in particular, that there are "fermionic" terms that complete the theory. We omit the fermionic completions for ease of notation and exposition.

analytic continuation to work out the answer for ordinary time. The advantage is that the Euclidean exponential weighting becomes $\exp(-S_E/\hbar)$, so the minima of S_E receive the greatest weighting and the integral might converge. The nonconstant minima of the Euclidean action are called *instantons*. After Euclideanizing equation (4), the action becomes the “energy” S_E of the map φ :

$$S_E = \int_{\Sigma} |\nabla \varphi|^2.$$

The energy of a map has a *conformal symmetry*, meaning that it is independent of local scale transformations on the Riemann surface, that is, transformations that can be locally approximated by a combination of rotations and dilations. Invariance under rescaling by a positive number λ can easily be seen: each of the two derivatives in $|\nabla \varphi|^2$ decreases by a factor of λ , while the area element increases by λ^2 . Rotational invariance is clear from the form of $|\nabla \varphi|^2$. The combination of the two, along with the fact that this argument did not depend on the derivatives of the scaling parameter λ , leads to the statement of local scale invariance.

The conformal symmetry of the action is an example of a classical symmetry of the action that is not necessarily maintained in the quantum theory. However, the quantum theory has no anomaly—meaning that the symmetry is preserved—if M is chosen to be a complex, CALABI-YAU MANIFOLD [III.6].

The Calabi-Yau condition can be thought of as a complex notion of orientation. Recall that for an oriented manifold one can continuously choose, on each patch, a basis for the tangent space such that, when we move from patch to patch, the determinant of the change-of-basis matrix is equal to one. The same is true on a Calabi-Yau manifold, but now we consider complex bases for the complex tangent spaces.

When the target manifold is a Calabi-Yau manifold, the instantons are complex analytic maps from the two-dimensional surface. Instantons are not “close” to the constant paths; their effects are therefore not accessible by perturbative methods such as Feynman diagrams. They are therefore “nonperturbative” phenomena. An example from quantum mechanics would be a particle in a double-well potential such as $(x^2 - 1)^2$. The zero-energy minima are the two constant (stationary) paths at $x = \pm 1$. An instanton path could go from $x = -1$ to $x = +1$, or vice versa. Such trajectories occur and are known as “quantum tunneling.”

Lesson. Inaccessible by perturbation theory, instantonic effects are notoriously challenging to calculate.

3.1 Mirror Pairs

In the setting above, we considered maps from a two-dimensional surface Σ to a target (Calabi-Yau) space. Let us denote this quantum field theory by $Q(M)$, which is shorthand for the collection of all fields and all possible correlation functions created from them. In this setup, we say that the Calabi-Yau manifolds M and W are “mirror pairs” if $Q(M)$ is equivalent to $Q(W)$. Through the magic of mirror symmetry, hard problems in $Q(M)$ involving instantons can be answered in $Q(W)$ by considering only the much simpler constant paths.

4 Mathematical Distillation

A physical theory contains a tremendous amount of information. For example, correlation functions can involve any number of fields, each evaluated at different points on the two-dimensional surface. This is typically too unwieldy a situation to approach mathematically. Instead, equipped with a symmetry of the theory called “supersymmetry,” a mathematical distillation can be performed. The distillation procedure is called *topological twisting*, and the resulting “topological field theory” has correlation functions that are independent of the positions of points. Because of this independence, the correlation functions are certain characteristic numbers associated with the underlying geometric setup. In fact, there are two types of twisting, typically called A and B, which capture different aspects of the manifold in question.

4.1 Complex and Symplectic Geometry

4.1.1 Complex Geometry

To get a feel for the geometric aspect captured by topological twisting, recall that we can construct the circle S^1 from the real line \mathbb{R} by identifying the points θ and $\theta + 2\pi$, and therefore also $\theta + 2\pi n$, where n is any integer. What we have done is identified points related by a *lattice of integer translations*. We could choose the lattice to consist of multiples of some other real number r , but since any two such lattices differ only by an overall scaling of \mathbb{R} , we would effectively get the same space. In the complex plane \mathbb{C} , we can do the same thing with a two-dimensional lattice of translations generated by two complex numbers λ_1 and λ_2 , as long as the quotient λ_2/λ_1 is not real. This space is called a *torus* and

has the same topology as any two-dimensional surface with one hole. It has more structure, however, because it can be covered by regions described by a complex coordinate—with different regions related by complex analytic maps. The pairs (λ_1, λ_2) and $(\lambda_1, \lambda_2 + \lambda_1)$ generate the same lattice of translations, as do the pairs (λ_1, λ_2) and $(\lambda_2, -\lambda_1)$. In fact, lattices related by a complex rescaling of \mathbb{C} are equivalent, so a better parametrization of the lattice is the ratio $\tau = \lambda_2/\lambda_1$.

By redefining the direction of one of the λ s, we can assume that the imaginary part of τ is positive, so τ takes values in the upper half of the complex plane. By the reasoning above, we note that τ and $\tau + 1$, as well as $-1/\tau$, all come from the same lattice. The number τ can also be thought of in the following way. The torus has two distinct loops, one generated by a straight path from z to $z + \lambda_1$, and one generated by a straight path from z to $z + \lambda_2$. Then λ_1 and λ_2 are both the result of the line integral of the complex differential dz over the loop. In fact, the loop did not even need to be straight to lead to this conclusion. The values of such integrals over subspaces without boundaries (the loops, here) are more generally called *periods*.

Although any two tori are topologically equivalent, one can show that there is no *complex analytic* map between two complex tori described by genuinely different values of τ . The parameter τ therefore determines the complex geometry of the space. Roughly speaking, we think of this parameter as describing the *shape* of the torus. (See MODULI SPACES [IV.8 §2.1] for a further discussion of this.)

The topological B-model depends only on the complex geometry of the target space M . That is, the theory depends, continuously, only on the parameter τ .

4.1.2 Symplectic Geometry

Another aspect of geometry is the *size* of the torus, which is described simply by an area element. Let us recall that, topologically, all tori look like \mathbb{R}^2 with points identified by the lattice of integer horizontal and vertical translations (but not necessarily in a way that would respect any complex geometry). The points of the torus can be thought of as the unit square with opposite sides glued together. An area element in \mathbb{R}^2 looks like $\rho \, dx \, dy$, which then determines the area ρ of the unit square. These notions of two-dimensional area generalize to two-dimensional subspaces in higher-dimensional spaces. The study of such structures is called SYMPLECTIC GEOMETRY [III.90], and so we call ρ the *symplectic parameter*.

The topological A-model depends only on the symplectic geometry of the target space M . That is, the theory depends, continuously, only on the parameter ρ .

4.2 Cohomological Theories

As you might imagine, the passage from an ordinary theory to a topological theory involves identifying many aspects of the physical theory that were previously distinct, such as different point values of a single field. Mathematically, a well-established method of producing topological aspects of a structure—and one that involves making identifications—is through a COHOMOLOGY THEORY [IV.6 §4]. Cohomology theories follow the pattern of having an operator δ obeying the equation $\delta \circ \delta = 0$. We think of this equation as the statement $\ker(\delta) \subset \text{image}(\delta)$. The cohomology group $H(\delta)$ is formed as the quotient $H(\delta) = \ker(\delta)/\text{image}(\delta)$, which means that we identify any two vectors u and v satisfying $\delta u = \delta v = 0$, so long as the difference $u - v$ can be written as δw for some w . Then $H(\delta)$ is just the space of all such vectors, up to identifications.

The topological twisting of physical theories is similar. The operator δ is a physical operator acting on a Hilbert space of states. The presence of supersymmetry in our theories ensures that δ exists and squares to zero. The vector states of the topological theory are just the elements of $H(\delta)$, i.e., states in the original theory Ψ obeying $\delta\Psi = 0$, up to identification. In many cases, these states can be identified with ground states.

It is crucial that supersymmetry is a symmetry that contains the complex translations of points on the two-dimensional surface. This means that the value of a field operator $\phi(z)$ at one point is identified with its value $\phi(z')$ at another. In other words, the physics of the topological theory is independent of the positions of the operators! In the path-integral formulation, this means that the correlation functions are independent of the positions of the fields inserted into the integrand. What can they depend on, then? They depend on the particular field or combination of fields inserted, and they depend on the geometric parameter (such as ρ or τ) of the space M .

4.2.1 The A-Model and the B-Model

Given a Calabi-Yau space, one can actually construct two operators, δ_A and δ_B , each of which squares to zero. There are therefore two distinct corresponding topological twistings and two distinct topological theories that can be constructed from a Calabi-Yau space.

If M and W are mirror Calabi-Yau pairs, you might wonder if the topological models constructed from them will still be equivalent theories. The answer is a most interesting form of yes: the resulting A-model of one Calabi-Yau manifold M is equivalent to the B-model of the mirror W , and vice versa! The complex and symplectic aspects of the theories get interchanged under mirror symmetry! In particular, a hard symplectic question of M might get mapped to an easy computation involving the complex geometry of W .

We emphasize here that the two manifolds may be completely topologically distinct. For example, the Euler characteristic of one is the negative of the other.

5 Basic Example: T-Duality

Although the circle is not complex, it provides a very illustrative entry into mirror symmetry that can be studied quite easily. We will find an equivalence between two theories constructed from circles. The equivalence will be very nontrivial, however, as states of very different kinds will be shown to correspond.

Consider the case where the two-dimensional surface is a cylinder, with spatial dimension a unit circle, and one dimension of time, and let us look at the sigma model (these were introduced in section 3). Suppose also that the target space is a circle of radius R , which we denote by S_R^1 . We think of S_R^1 as the real line, with two points identified if they differ by a multiple of $2\pi R$. Maps from one circle to another can be classified by their *winding number*, an integer that tells you how many (net) times the image of a point goes around the second circle when the point goes once around the first. The map $\theta \mapsto mR\theta$ from the circle to S_R^1 has winding number m . This allows us to write the field $\varphi(\theta)$ as a winding piece, $mR\theta$, plus an honest Fourier series (no winding): $\varphi(\theta) = mR\theta + x + \sum_{n \neq 0} c_n \exp(in\theta)$. Here we have singled out the constant mode $x = c_0$ of the Fourier series. We have expanded just the θ dependence in a series, so every continuous parameter (x and the c_n) should be thought of as a function of time, as well.

The energy, or Hamiltonian, of such a map is computed as in section 2.1.3:

$$H = (mR)^2 + \dot{x}^2 + \sum_n |\dot{c}_n|^2 + n^2 |c_n|^2.$$

Comparing this with the harmonic oscillator Hamiltonian of section 2.1.3, we can see that each degree of freedom $c_n(t)$ plays the role of a (complex) quantum-mechanical particle in a simple harmonic oscillator potential. There is an occupation-mode basis for

describing the quantum mechanics of each mode.⁹ The full Hilbert space of the quantum theory is the (tensor) product of each of these, plus parts involving the constant mode and winding number, which we now discuss. (Remember, each degree of freedom of the classical theory becomes a *particle* in the quantum field theory.)

The constant mode x has energy \dot{x}^2 , and therefore has no associated potential (it can be anywhere on the circle). This mode represents a free quantum-mechanical particle on the circle. Recall that the momentum of the x particle is represented by the operator $-i(d/dx)$. This operator has eigenfunctions e^{ipx} . The requirement that these eigenfunctions are invariant under the translation $x \rightarrow x + 2\pi R$ means that the eigenvalues of momentum are “quantized,” and have the form $p = n/R$.

In contrast to momentum, the integer winding number (m) is really a classical label for the possible maps from a circle to a circle. Although integral, it is clearly on a different footing from the integer n of momentum. Still, it is also an important label on the Hilbert space. For each m , we have a space of m -winding configurations which gets quantized to become the m th sector of the Hilbert space. Roughly, this sector \mathcal{H}_m comprises the functions of all the degrees of freedom of all the m -winding maps. We can consider the winding number as an operator by simply declaring that the states with winding number m have eigenvalue mR .

Ignoring the oscillator modes for the moment, the state of momentum n/R with winding m has energy $(n/R)^2 + (mR)^2$. In particular, the energy is unchanged if we make the simultaneous switches $(m, n) \leftrightarrow (n, m)$ and $R \leftrightarrow 1/R$. Since the oscillator modes a_n have energies that are independent of R , and since the modes are noninteracting particles, this symmetry can be extended to a full equivalence of the theories with targets S_R^1 and $S_{1/R}^1$, with momentum in one theory corresponding to winding number in the other.

In this example, the target space S^1 is neither complex nor symplectic. As a result, we cannot construct the topological A- and B-models. Nevertheless, we have demonstrated the stronger statement that the two sigma models with target space S_R^1 and $S_{1/R}^1$ are equivalent. The theories are mirror pairs. In the special case of circles, mirror symmetry is referred to as T-duality. In fact, the entire phenomenon of mirror symmetry—even for noncircles—can be deduced from T-duality.

9. Each $a_n^\dagger = [\text{Re}(\dot{c}_n) - i n \text{Re}(c_n)]/\sqrt{2n}$ is a raising operator, and similarly for the imaginary parts of the c_n .

5.1 Tori

If we take the product of two circles $S^1_{R_1} \times S^1_{R_2}$, we get a torus. We can think of the torus as a circle family of circles, since for each point in $S^1_{R_2}$ we have a circle $S^1_{R_1}$. As we have seen in section 4.1.1, this space is complex—specifically, it is the complex plane \mathbb{C} quotiented by a lattice of translations. A particularly simple lattice is the one generated by the translations $z \rightarrow z + R_1$ and $z \rightarrow z + iR_2$. As discussed in section 4.1.1 above, the lattice is determined by the complex number $\tau = iR_2/R_1$, equal to the ratio of integrals (“periods”) of the complex form dz over the two nontrivial loops of the torus.

The symplectic data is captured by the area element. Recall that we can choose coordinates x and y such that the identifications look like unit translations in each direction. Then the (normalized) area element of the torus with radii R_1 and R_2 is $R_1 R_2 dx dy$, which integrates to $R_1 R_2$ on the unit square. Let us define the symplectic parameter $\rho = iR_1 R_2$. We now perform T-duality for the first circle $R_1 \rightarrow 1/R_1$. We see that under this substitution, the complex and symplectic parameters get interchanged:¹⁰

$$\tau \longleftrightarrow \rho.$$

Lessons. Mirror symmetry interchanges complex and symplectic parameters. Mirror symmetry is T-duality.

5.2 The General Case

The torus is the only compact one-dimensional Calabi-Yau space and is therefore the simplest one, but the discussion above is part of a more general picture. The Calabi-Yau condition ensures a unique complex volume element, or orientation (dz , above), whose “periods” determine, and in turn vary with, the complex parameters. Though the A- and B-models both turn out to be rather simple in the case of the torus, what is important in general is that the B-model is completely determined by how the periods of the complex volume element (which were λ_1 and λ_2 in section 4.1.1) change with the parameters of the theory (of which there was just one in section 4.1.1, namely τ). Again, the relation $\tau = \lambda_2/\lambda_1$ is quite simple for the torus, but more complicated in general. In any case, this data gives all the information of the B-model. The reason for all of this is that the instantons of the B-model turn out to be just the constant maps. Each point of the target space determines a constant map, and as a result the B-model is

reduced to (classical) complex geometry of the target space. This is determined by the periods.

This state of affairs is to be compared with the A-model. The A-model depends on the symplectic parameters ρ , i.e., the areas of two-dimensional surfaces inside the target space. In contrast to the B-model, however, the dependence on ρ is very complicated, in general. The reason for this is that the instantons of the A-model are area-minimizing surfaces inside the target space, and their enumeration is a notoriously challenging problem. (The problem is not terribly challenging for the torus, however.) Mathematically, the A-model instantons are described by the theory of Gromov-Witten invariants, the subject to which we now turn.

6 Mirror Symmetry and Gromov-Witten Theory

As we mentioned above, the B-model on W is explained entirely by the classical complex geometry of W . The only relevant maps for B-model computations are the constant ones, so the space of such maps is equal to W itself, and correlators reduce to (classical) integrals over W . In fact, one of the integrands to be integrated is the complex volume element. Let us call the parameter for all possible complex volume elements τ . B-model correlation functions are then determined by τ -dependent integrals over W . In particular, the partition function $Z_B^{(W)}$ of the B-model on W depends on τ , so we write it as $Z_B^{(W)}(\tau)$.

The main point about topological twisting is that local variations of the fields are all identified, as they are related by the operator δ . In particular, varying the point on the worldsheet is a trivial operation in the topological theory. It turns out that, for the B-model on W , *only* the constant maps contributed, but for the A-model the situation is a bit more subtle. To give a feel for the geometry, consider again the winding of a map from a circle to a circle. Maps with different windings can never be deformed continuously into one another. The winding number is a measure of how the first circle “wraps” (or winds) around the target, according to the map. Because it is a discrete parameter it cannot change under continuous variations. Likewise, when M is a higher-dimensional space, the two-dimensional surface Σ can “wrap” around two-dimensional subspaces of M by different amounts. The parameters for wrapping are again discrete. A map φ can wrap Σ around the basic surfaces C_i in M by different integer amounts,

10. The parameters τ and ρ can also have real parts, but we neglect the details for simplicity.

k_i . We say that $k = k_i$ labels the “class” of the map φ . (More precisely, $\varphi(\Sigma)$ is a closed 2-cycle when Σ is compact, and k labels its homology class.) Different classes k contribute through different (Euclidean) actions $S_k(\rho)$, which depend on the areas ρ and the class k but not on the continuous details of the map φ_k . The partition function can have contributions from all classes. Different classes may contribute differently not only through the exponential weighting, but also in accordance with how many *minimal surfaces* they contain. (A good example of a minimal surface in three-dimensional space is a soap film. If you fix the boundary with a wire, the soap film will seek to find the minimum-area surface with that boundary.) In our examples, the space M is actually complex; the minimal surfaces we speak of in Gromov–Witten theory are complex analytic maps from Σ . That is, if you have a complex coordinate for Σ , then the complex coordinates for the surfaces M can be written as complex analytic functions of Σ .

The difference between the A-model and the B-model comes from the fact that the topological model is constructed from an operator δ , which was guaranteed to exist by the presence of supersymmetry in our theories. For the different models, the relevant supersymmetry operators δ_A and δ_B are simply different. As we saw above, the maps relevant to the A-model are the instantons, or complex analytic maps from Σ to M . Roughly, then, A-model correlation functions on M , and in particular the partition function $Z_A^{(M)}$, are sums over classes k of surfaces in M and sums over instantons in each class, each one weighted by its instanton action $\exp(-S_k(\rho))$. We have explicitly written the dependence on the parameter for the symplectic structure ρ . For Calabi–Yau manifolds, such maps should be discrete, and it is a conjecture, true in all known cases, that they are finite in number if we fix the class, k . All this data is packaged in a function of ρ , and based on what we have argued, the partition function must take the general form

$$Z_A^{(M)}(\rho) = \sum_k n_k \exp(-S_k(\rho)).$$

The coefficients n_k are called *Gromov–Witten invariants*.¹¹

Putting things together, if (M, A) is mirror to (W, B) , and if we can identify for each complex parameter τ

for W a corresponding symplectic parameter $\rho(\tau)$ for M , then we have

$$Z_A^{(M)}(\rho) = Z_A^{(M)}(\rho(\tau)) = Z_B^{(W)}(\tau). \quad (6)$$

The first equality means we should rewrite ρ in terms of τ , and the second says that the answer should be given by the corresponding B-model on W . Therefore, all of the information about complex analytic surfaces in M , which is encapsulated in the coefficients n_k , is completely determined by the classical geometry of W !

This remarkable predictive power—the computation of an infinite number of difficult Gromov–Witten invariants through equations such as (6)—is what led to such intense interest in mirror symmetry at its inception.

7 Orbifolds and Nongeometric Phases

7.1 Nongeometric Theories

Mirror symmetry is about an equivalence of quantum field theories, and not every such field theory has the geometric content of a target space as in the sigma model. The structure involved in mirror symmetry—or at least its topological version—begins with a quantum theory with a supersymmetry algebra that allows for the passage to a topological theory. That is, there is a Hilbert space of states, a Hamiltonian operator, and a particular algebra of symmetries, i.e., operators that commute with the Hamiltonian. There are no dictates as to how one constructs such a setup, and the sigma model of maps to a target space is only one such way. Other methods abound. The geometric case is merely the one most suited for mathematicization (and exposition), which is why we have focused on the theory with a target space.

As an intermediate case—possibly geometric, possibly not—we will discuss the so-called orbifold theories.

7.2 Orbifolds

When space-time is a cylinder $S^1 \times \mathbb{R}$, with a circle S^1 as its spatial dimension, there is a fascinating construction in quantum field theory known as an *orbifold theory*. This is defined as follows. Suppose there is a finite group G of symmetries (such as a reflection symmetry). That is, each group element acts as an operator on the Hilbert space, so if $g \in G$ then it sends a state Ψ to a state $g\Psi$. Then one defines a new theory by identifying states related by the symmetry. To construct the theory, let us first consider the ground state Ψ_0 of the original theory. This is assumed to be invariant under the group: that is, $g\Psi_0 = \Psi_0$ for all group

11. Though our discussion makes it seem as though the n_k are integers, in fact they are only rational numbers. They can be expressed in terms of more basic integers, however. These integers are the ones referred to at the beginning of this article.

elements, g .¹² One then constructs the space \mathcal{H}_0 of all invariant states. This is known as the *untwisted sector*, and Ψ_0 is the ground state of the untwisted sector. In the case where G is commutative, a *twisted sector* is then constructed for every group element $g \in G$.¹³ To construct the twisted sector, first think of the spatial dimension S^1 as being an interval $[0, 1]$ with endpoints 0 and 1 identified. Recall that the Hilbert space of states is constructed from (functions of) all the degrees of freedom of the possible configurations of fields. The twisted sector \mathcal{H}_g corresponds to additional field configurations Φ that are related at the two ends by the action of g : so $\Phi(1) = g\Phi(0)$. Such field configurations represent configurations on the circle S^1 since left and right ends are related by the group, and therefore get identified. These additional configurations are thus part of the orbifold theory. One constructs a sector \mathcal{H}_g of the Hilbert space by taking all such states Ψ_g that also obey the invariance condition $h\Psi_g = \Psi_g$ for all group elements h .

Orbifolds may be geometric, as they are in the case of the sigma model to a manifold X on which a discrete group G acts. For example, rotations act on the plane, and we can consider the four-element group generated by a right-angle rotation. The quotient of the plane by these rotations looks like a cone. As another example, the finite groups of symmetries of the platonic solids (tetrahedron, cube, etc.) act on the two-dimensional sphere by rotations. When we take $X = S^2$ and G a platonic group, we get an interesting orbifold. In fact, if we simply take the space of orbits of the group G , it is topologically just a sphere again, but not a smooth one—it has cone points. These cone points would be troublesome in a quantum field theory, but the “stringy” orbifold is perfectly “smooth.”

The orbifold theory itself carries a symmetry. For example, if G is the commutative group with two elements, then there is an untwisted sector and a unique twisted sector. There is a symmetry corresponding to multiplication by 1 in the untwisted sector and by -1 in the twisted sector. This symmetry is not geometric. Orbifold theories with symmetries can often themselves be orbifolded in such a way as to recover the original theory. In fact, the theory and its orbifold are

also often mirror pairs! Greene and Plesser used such a construction to create the first examples of mirror pairs. Furthermore, they used ways of ascribing geometric interpretations to some nongeometrically constructed theories so as to identify mirror Calabi-Yau spaces. To be precise, they took the space of all nonzero complex 5-vectors $X = (X_1, X_2, X_3, X_4, X_5)$ satisfying the equation

$$X_1^5 + X_2^5 + X_3^5 + X_4^5 + X_5^5 + \tau X_1 X_2 X_3 X_4 X_5 = 0,$$

identifying X with λX for any nonzero complex number λ . (If X is a solution, then so is λX .) The equation actually defines a family of complex spaces, since $\tau \in \mathbb{C}$ is a parameter. The orbifold theory is defined from the finite group of phase transformations

$$(X_1, X_2, X_3, X_4, X_5) \mapsto (\omega^{n_1} X_1, \omega^{n_2} X_2, \omega^{n_3} X_3, \omega^{n_4} X_4, \omega^{n_5} X_5),$$

where $\omega = e^{2\pi i/5}$ and $\sum_{i=1}^5 n_i$ is a multiple of 5. This space and its orbifold are actually the mirror pair about which Candelas et al. made their famous predictions.

8 Boundaries and Categories

The entire story of mirror symmetry becomes much richer when we allow the strings to have endpoints. Strings with ends are called “open strings,” while “closed strings” refers to loops. Mathematically, allowing ends corresponds to adding boundaries to the worldsheet surfaces. With this addition, we would like to perform the same topological twisting. To do so, we must first ensure that some supersymmetry condition persists when we put the boundary conditions on the fields. If we begin with a Calabi-Yau target manifold, we can ask to preserve the conditions that allow either the A-twisting or the B-twisting (but not both: the boundary condition will destroy some symmetry, much as pinning a rope will constrain its degrees of freedom). After the twist, the boundary topological theory will depend on symplectic or complex information, respectively.

For the A-model, the endpoints or boundaries must lie on a Lagrangian subspace. The Lagrangian condition constrains half the coordinates; for linear spaces it is like a restriction to the real part of a complex vector space. For the B-model the boundaries must lie on a complex space. Locally, a complex space looks like \mathbb{C}^n and a complex subspace is described by complex analytic equations in the coordinates. A boundary condition that preserves supersymmetry and allows a chosen topological twisting is called a *brane*. (The terminology mimics the word “membrane,” but applies to any

12. In the case where there are flat directions of a potential, as in a free particle on a circle (no potential at all), the ground state may be a superposition of classical values of the field. For the circle, the constant wave function $\Psi = 1$ is not associated with a single, classical location. It is still invariant under any group of rotations, however.

13. The twisted sectors are properly labeled by conjugacy classes, which are the same as group elements when G is commutative.

dimension.) In short, A-branes are Lagrangian; B-branes are complex.

To package all the information of the topological boundary theory, one appeals to the mathematical notion of a CATEGORY [III.8]. A category is a way of talking about structure: it consists of *objects*, and for any pair of objects there is a space of *morphisms* from one object to the other. Often the objects are mathematical structures of some kind and the morphisms from one object to another are the functions that preserve the relevant structure. For example, if the objects are (i) SETS [I.3 §2.1], (ii) TOPOLOGICAL SPACES [III.92], (iii) GROUPS [I.3 §2.1], (iv) VECTOR SPACES [I.3 §2.3], or (v) chain complexes, then the morphisms are, respectively, (i) MAPS [I.2 §2.2], (ii) CONTINUOUS MAPS [III.92], (iii) HOMOMORPHISMS [I.3 §4.1], (iv) LINEAR MAPS [I.3 §4.2], or (v) chain maps. The morphism spaces between objects should be thought of as some kind of relational data. Morphisms themselves interact with one another, as they can be composed when the end object of one morphism is the start object of another. The composition is associative, so whether you compute abc as $(ab)c$ or $a(bc)$ does not matter. A useful image is a directed graph, which is a category with vertices as objects and paths between two vertices as morphisms. Composition is defined in this category by concatenating paths.

In the case of a two-dimensional field theory with boundary conditions, we construct a category whose objects are branes (i.e., boundary conditions). The morphisms between two branes α and β are the ground states $\mathcal{H}_{\alpha\beta}$ of the boundary field theory defined on the infinite strip $[0, 1] \times \mathbb{R}$, where we put the boundary condition α on the left boundary $\{0\} \times \mathbb{R}$ and the condition β on the right boundary $\{1\} \times \mathbb{R}$. Morphisms are composed by gluing boundaries together, and associativity is guaranteed by topological invariance.¹⁴

Mirror symmetry with boundary conditions then becomes the following statement: two manifolds M and W are mirror pairs if the brane category of the A-twisting of M is equivalent to the brane category of the B-twisting of W (and vice versa). The mathematical translation of this statement is called the *homological mir-*

ror symmetry conjecture, due to Kontsevich. On the A-model side, the brane category is the so-called *Fukaya category*, and is governed by complex analytic maps from surfaces with boundaries, where the boundaries must be mapped to Lagrangian branes. On the B-model side, the branes form a category determined by complex subspaces, together with complex analytic VECTOR BUNDLES [IV.6 §5] on them. A complex vector bundle associates a complex vector space to every point. For example, the complex circle $\{x^2 + y^2 = 1\}$ in \mathbb{C}^2 has a complex tangent space at every point. “Complex analytic” means that this subspace of \mathbb{C}^2 changes in a complex analytic way. For the complex circle, the space of tangent vectors at (x, y) consists of all multiples of the vector $(-y, x)$, an assignment which is clearly complex analytic. Physically, the bundles arise from allowing charges on the endpoints of strings.

Kontsevich’s conjecture asserts that these two categories of branes are equivalent. That statement is natural from the physics point of view, but by identifying the precise categories that correspond to the physical picture, this conjecture is a major contribution to the translation of mirror symmetry from physics into rigorous mathematics. The equivalence of categories means that not only is there a corresponding Lagrangian A-brane of M for every complex B-brane of W , but that the *relationships*, or morphisms, between branes are also in correspondence.

8.1 Example: Torus

Kontsevich’s conjecture can be proven and easily illustrated in the example of a 2-torus. Think of the now-familiar symplectic two-torus as being the two-dimensional plane, with integer lattice translations identified. We take the torus to have area element $A dx dy$, so that the symplectic parameter is the imaginary number $\rho = iA$, as in section 4.1.2. Now consider straight lines on the plane. These will correspond to closed circles on the torus as long as they have rational slope: $m = d/r$, with d and r relatively prime integers. They are Lagrangian branes of the A-model boundary theory. The minimal-energy open strings connecting one line of slope $m = d/r$ to another of slope $m' = d'/r'$ are those that have zero length. They are therefore the points of intersection. It is an easy exercise to show that there are $|dr' - rd'|$ such points.

On the mirror side, we again have a torus, but with a complex parameter τ , and for the two tori to be mirror pairs, we should set $\tau = \rho$. The objects of the B-model

14. We speak of associativity of the topological states, which are themselves cohomology classes. At the “chain” level, before the topological twisting, there is no associativity. The notion of a category with morphisms that have a cohomology and compose only “up to cohomology” is called an A_∞ category. One can also imagine a categorical definition that captures the structure of surfaces with handles and holes. Indeed, the proper mathematical framework for a complete understanding of mirror symmetry is still under construction.

brane category are complex vector bundles. It is a theorem that the basic bundles are classified by their rank r and degree d , two integers.¹⁵ It is customary to organize these two numbers into what is known as a “slope,” $m = d/r$ (the nomenclature preceded this application), and basic bundles must have d and r relatively prime.

We can now easily guess that under the mirror correspondence we have

$$\text{slope} \longleftrightarrow \text{slope}.$$

This means that a Lagrangian brane of slope m on the torus with symplectic parameter ρ should correspond to a complex vector bundle with slope m in the mirror torus with complex parameter ρ . Now suppose we have the B-model version of our example above, so we take two vector bundles of slope m and m' . In fact, the minimum-energy open strings between two complex analytic bundles of slope m and m' correspond to complex maps between the bundles, and the RIEMANN-ROCH FORMULA [V.34] counts this number as $|dr' - rd'|$. This is the same result as for our A-model calculation above! Therefore, corresponding objects relate in a corresponding way. Beyond the morphism spaces, one checks finally that the compositions of corresponding morphisms correspond, just as for logarithms and slide rules. Doing so proves Kontsevich’s conjecture.

8.2 Definition and Conjecture

In fact, Kontsevich’s definition of mirror symmetry is really a conjecture stating that the boundary notion of mirror symmetry as an equivalence of categories is compatible with, and even implies, the traditional notion of mirror symmetry that relates Gromov–Witten theory and complex structures.

One way to show this is to try to reconstruct the Gromov–Witten invariants from the boundary theory. A heuristic, geometric approach to doing so involves looking at the diagonal boundary condition in two copies of a space. A disk mapping into two copies of a space is described by two maps of a disk into the space. Further, if the boundary condition is diagonal, this means that the maps have to agree on the boundary. What we have, then, is two disks inside a space which agree on the boundary. That is exactly what a sphere is: two disks (or cups) glued together!

15. A vector bundle assigns a vector space to each point of the torus. The rank is the dimension of that space. The degree is roughly a measure of the complexity of the bundle. For example, if we have a two-dimensional surface and consider the bundle that assigns to each point the tangent space at that point, the degree is equal to $2 - 2g$, where g is the number of holes on the surface.

The disks are the two hemispheres, and they are glued along the equator. Now the minimal disks are instantons for the open string (with boundary), and by gluing them together along a common boundary, we have constructed a minimal sphere, or closed-string instanton. Thus the open string on this double theory should recover the closed string on the original theory.

A more algebraic approach sees the closed-string deformations as deformations of the category of branes. That is, a change in bulk (nonboundary) theory induces a change in boundary theory. But once equipped with a category, one can classify its deformations intrinsically. That is, if one views a category as a fancy algebra,¹⁶ then, as the deformations of an algebra are easily classified through a notion called Hochschild cohomology, the deformations of a category can be treated similarly. One arrives at the maxim that the closed string is the Hochschild cohomology of the open string. By computing the Hochschild cohomology of a brane category, one can, in principle, check this maxim, establish Kontsevich’s conjecture, and then prove the connection to traditional mirror symmetry and Gromov–Witten theory.

9 Unifying Themes

How does one find mirror pairs (M, W) ? What is the construction? Although mirror symmetry has spawned many results and proofs, these basic questions continue to vex.

On the one hand, Hori and Vafa have given a physics proof of mirror symmetry, which constructs mirror pairs but not through an evident mathematical channel. Of course, one can mathematicize the physical argument, but that does not seem to lead to insights into the construction—perhaps because path integrals and other methods of quantum field theory such as renormalization are not very well understood mathematically.

Batyrev has devised a procedure for constructing mirror pairs within the context of toric geometry. This method is a generalization, to a wide class of examples, of the original construction of Greene and Plesser. The recipe has been extremely successful in producing examples of every stripe. However, the underlying meaning behind the construction is unclear.

As for a geometric construction of mirror pairs, there is a physical argument that makes contact with mathematics, but it has not yet been made rigorous. The

16. An algebra is a category with one object.

argument uses T-duality. Start with the B-model on M and consider a point P of M as a zero-dimensional complex subspace. Then the choice of point P on M is parametrized by M itself. By mirror symmetry, there should be a corresponding Lagrangian brane T on the mirror manifold W . Furthermore, the choices of T must equal the choices of P , i.e., the manifold M . Therefore, if we can find the brane T on W , we can parametrize the choices of T , and recover M . So we can find the mirror M of W from W itself.

This construction is geometric and has something to say about the structure of the Calabi-Yau spaces involved in mirror symmetry. Specifically, the choices of a Lagrangian brane always look like a family of tori. Therefore, M itself should look like a family of tori. Further, one can argue that by performing T-duality in families of tori (in a similar way to how one does it for a single torus), one arrives back at the mirror manifold, W . This is what we did for the torus, thought of as a circle family $(S^1_{R_2})$ of circles $S^1_{R_1}$. When we T-dualized each member of the family, we found the mirror torus. So mirror symmetry is T-duality, and Calabi-Yau spaces of mirror symmetry should look like families of tori. This approach also relates to the homological mirror symmetry construction. Though promising, it remains mathematically elusive.

Various points of view on mirror symmetry are helpful for different applications. To date, no unified understanding of the phenomenon has been achieved. To some extent, we are still “feeling the elephant.”

10 Applications to Physics and Mathematics

As a computational tool in string theory, mirror symmetry is unparalleled in its power. When combined with other physical equivalences, its power is multiplied. For example, there are certain equivalences in physics that relate one type of string theory to another.

Without going into the details of string theory, we can get a flavor of its complexity by returning to mirror symmetry. Recall that the B-model was able to compute the difficult instantons on the A-model, yielding a great simplification of the two-dimensional quantum field theory on the worldsheet. But this whole quantum field theory was just an *auxiliary* tool for computing some Feynman diagram for the perturbation theory of the full string theory! Unfortunately, a satisfactory description of the full string theory path integral is, at the time of writing, way out of reach. String theory instanton effects are mostly unknown to us, unless a string

equivalence or other argument can relate them to a perturbative effect in a *different* string theory. The perturbative string calculation in that other theory may then be performed by exploiting mirror symmetry. Tracing through chains of equivalences in such a manner, many different phenomena in string theory can ultimately be calculated via mirror symmetry.

In principle, one should be able to calculate *all* non-perturbative and perturbative aspects of a single theory by outsourcing the calculations to equivalent theories and exploiting mirror symmetry. The barriers to doing this at the time of writing are largely technological, not conceptual.

Beyond physics, the rich texture of mirror symmetry means that there is interesting mathematics to be discovered in the proper formulation of the problem. For example, defining the precise categories of branes in full generality remains a challenge.

Yet there are also direct applications to mathematical questions. We have already discussed how enumerative geometry has been revolutionized by mirror symmetry and the counting of instantons. Results in symplectic geometry have also been obtained. Occasionally, two objects may be proven to be equivalent as B-model branes. If the A-model mirrors can then be found, one has the result that the corresponding Lagrangian subspaces of the mirror symplectic space are also equivalent. Of course, to make such an argument, one must first prove Kontsevich’s version of mirror symmetry for the mirror pair considered. As a final recent example, Kapustin and Witten have found a relation of mirror symmetry to the geometric Langlands program in representation theory. This program, loosely stated, is a correspondence between objects associated with two-dimensional surfaces and Lie groups. From a surface Σ and a gauge group G , one constructs the space \mathcal{M}_H of solutions to Hitchin’s equations. Central to that program are complex analytic objects on \mathcal{M}_H that behave nicely under the action of an algebra of operations. The Langlands correspondence relates two sets of such objects: one easy to calculate and the other more difficult. In fact \mathcal{M}_H is itself a family of tori, and the easy objects correspond to points. Mirror symmetry states that the points should turn into the tori under T-duality, so the hard objects should correspond to the tori themselves! It is an appealing proposition, and making it precise mathematics will be difficult—but the gauntlet has been thrown down.

The discovery that mirror symmetry relates to the geometric Langlands program has elicited great excite-

ment among researchers and reveals yet another facet of this fascinating phenomenon.

Further Reading

The article “Physmatics” (which can be found online at www.claymath.org/library/senior_scholars/zaslow_physmatics.pdf) is a general discussion of the relationship between mathematics and physics, and may serve as a complement to this article. Readers with a university-level mathematics background who want to learn about mirror symmetry in more detail could try consulting the book *Mirror Symmetry* (Clay Mathematics Monographs, volume 1, edited by K. Hori and others (American Mathematical Society, Providence, RI, 2003)).

IV.17 Vertex Operator Algebras

Terry Gannon

1 Introduction

Algebra is the mathematics that places more emphasis on abstract structure than on intrinsic meaning. The conceptual simplifications that can result when context is stripped away from structure give algebra a special power and clarity compared with other areas: compare, for example, the difficulty of visualizing four-dimensional space with the triviality of manipulating quadruples (x_1, x_2, x_3, x_4) of real numbers. However, this abstractness can also blind us. For instance, basic identities like $ab = ba$ and $a(bc) = (ab)c$ that are obeyed by numbers can be modified in countless directions, and each modification defines a new algebraic structure, but it is hard to guess from a purely abstract perspective which of these modifications will give rise to a rich, accessible, and interesting theory. For guidance, algebra has traditionally turned to geometry. For example, over a century ago LIE [VI.53] suggested that the identities $ab = -ba$ and $a(bc) = (ab)c + b(ac)$ were worth studying for geometrical reasons: the resulting structures are now called LIE ALGEBRAS [III.50 §2]. More recently, as we shall see, physics has joined geometry in this guiding role and has had spectacular success.

The renowned physicist and mathematician Edward Witten believes that a major theme of twenty-first-century mathematics will be its reconciliation with the branch of physics known as quantum field theory. Conformal field theory (the quantum field theory that underlies string theory) is an especially symmetric and well-behaved class of quantum field theories. When this

notion is translated into algebra, the result is a structure known as a *vertex operator algebra* (VOA). This article sketches where VOAs come from, what they are, and what they are good for.

To aim to explain a VOA in a few pages is almost as absurd as to aim to explain quantum field theory in a few pages, but, undaunted, I shall try to do both. Obviously it will be necessary to gloss over many important technicalities and to commit major simplifications; without question this exposition will raise the ire of experts and the eyebrows of knowledgeable amateurs, but I hope that it will at least convey the essence of this important and beautiful area. Vertex operator algebras are the algebra of string theory: they should be thought of as the same sort of gift to the twenty-first century that Lie algebras were to the twentieth.

2 Where VOAs Come From

The two most revolutionary developments in physics in the early twentieth century are usually held to be relativity and quantum mechanics. They are revolutionary not just because they have consequences that are extremely counterintuitive, but also because they provide very general frameworks that can potentially affect all physical theories: one can take a theory from classical physics, such as the theory of the harmonic oscillator or the theory of electrostatic force, for example, and one can try to make it “relativistic,” so that it becomes compatible with relativity, or to “quantize” it, so that it becomes compatible with quantum mechanics.

Unfortunately, nobody knows how to make relativity fully compatible with quantum mechanics. To put this another way, the ultimate concern of relativity is gravitation, and a direct application to gravity of the usual quantizing techniques fails. This ought to mean that a fundamentally new physics arises at small distance scales that we are ignoring. Indeed, naive calculations suggest that the space-time “continuum” at distance scales of around 10^{-35} m should deteriorate into some sort of “quantum foam,” whatever that might mean. (10^{-35} m is extremely small: for instance, the order of magnitude of the size of an atom is 10^{-10} m.)

Perhaps the most popular and controversial approach to quantum gravity is string theory. The electron is a *particle*, i.e., in principle it can be localized to a point. In string theory, the fundamental object is a *string*, a finite curve of length approximately 10^{-35} m. In place of the dozens of kinds of fundamental particles in the generally accepted quantum field theory, there

is only one string, whose precise physical properties (mass, charge, etc.) depend on its current “vibrational mode.”

As the string moves, it traces out a surface called a worldsheet. For reasons that we will sketch below, much of string theory reduces to studying conformal field theory, which is the induced quantum theory on these surfaces. Probably no other structures have affected so many areas of “pure” mathematics in so short a time as string theory and, what is essentially the same thing, conformal field theory. Indeed, five of the twelve Fields Medals awarded in the 1990s (namely, those to Drinfel’d, Jones, Witten, Borcherds, and Kontsevich) were for such work. We shall focus in this article on their algebraic impact; see MIRROR SYMMETRY [IV.16] for some geometrical implications.

2.1 Physics 101

A quick overview of physics will be useful for the discussion. Further details can be found in MIRROR SYMMETRY [IV.16 §2].

2.1.1 States, Observables, and Symmetries

A physical theory is a set of laws that govern the behavior of some kind of physical system. A *state* of that system is a complete mathematical description of the system at a particular time: for instance, if the system consists of a single particle, then we could take its state to be its position x and momentum $p = m(dx/dt)$ (where m is its mass). An *observable* is a physically measurable quantity such as position, momentum, or energy. It is through observables that a theory is compared with experiment. Of course, for this to be true we also need to know what an observable is from a theoretical point of view. In classical physics, this is easy: an observable is just a numerical function of the state. For example, our single particle has energy E , which depends on the position and momentum via a formula of the form $E = (1/2m)p^2 + V(x)$. (This gives us the kinetic energy plus the potential energy.) Classical states at different times are related by the equations of motion, which are usually expressed as differential equations.

However, string theory and conformal field theory (CFT) are quantum theories, which are significantly different from classical theories: one can think of them as “applied linear algebra.” Whereas a classical state was given by a collection of a few numbers (two, in the case of the particle above), a quantum state is an

element of a HILBERT SPACE [III.37], which for the purposes of discussion we can think of as a column vector with infinitely many complex entries. As for a quantum observable, it is a HERMITIAN OPERATOR [III.52 §3.2] on the Hilbert space, which we can think of as an $\infty \times \infty$ matrix \hat{A} . This operator acts on the states by matrix multiplication. As in classical physics, one of the most important observables is energy, which is given by the *Hamiltonian operator* \hat{H} .

It is far from obvious how a linear operator that takes states to states has anything to do with the notion of a physical observation, and indeed the relationship between observables and observation is a major difference between classical and quantum theories. If \hat{A} is an observable, then the SPECTRAL THEOREM [III.52 §3.4] tells us that the Hilbert space has an ORTHONORMAL BASIS [III.37] of EIGENVECTORS [I.3 §4.3]. When we do the experiment that is modeled by the observable \hat{A} , the answer we obtain will be one of the eigenvalues of \hat{A} . However, this answer is usually not fully determined by the state v . Instead, it is given by a probability distribution: the probability of obtaining a particular eigenvalue is proportional to the square of the norm of the projection of v into the corresponding eigenspace. Thus, the only circumstances under which the answer is determined in advance are if the state v is an eigenvector of \hat{A} .

There are two independent ways in which a quantum state can evolve in time: a deterministic evolution between measurements, governed by the famous SCHRÖDINGER EQUATION [III.85], and a probabilistic and discontinuous one that occurs at the instant when a measurement is made. For our purposes, only the deterministic evolution will be relevant.

The symmetries of CFT are extremely rich, as we shall see. Symmetries in physical theories are highly desirable because of two consequences that they have. First, they lead by NOETHER’S THEOREM [IV.12 §4.1] to *conserved quantities*, i.e., quantities independent of time. For example, the equations of motion of our particles are usually invariant under translation: for instance, the gravitational force between two particles depends only on the difference between their positions. The corresponding conservation law in this case is the conservation of momentum. A second consequence of symmetries in quantum theories is that infinitesimal generators of the symmetries act on the state space \mathcal{H} (the Hilbert space to which the states belong), forming a representation of the Lie algebra. Both consequences are important to CFT.

PUP: this spelling is correct.

T&T note: Tim to think about all spectral theorem CRs later.

2.1.2 The Lagrangian Formulation and Feynman Diagrams

We will need two of the languages in which physics is written. One is the *Lagrangian* formalism, which is responsible for the relationship between string theory and CFT, as well as for the appearance of modular functions in string theory. The other is the *Hamiltonian* or *Poisson bracket* formalism, which is where algebra arises. Vertex operator algebras try to explain the “miracle” that these two formalisms cohere.

The Lagrangian formalism can be expressed classically through Hamilton’s *action principle*. When there are no forces present, particles travel in straight lines, which are the curves of shortest length. Hamilton’s principle explains how this idea generalizes to arbitrary forces: instead of minimizing length, the particle minimizes a related quantity S called the *action*.

The quantum version of Hamilton’s principle is due to Feynman. He expresses the probability of measuring the system in some final (eigen)state $|\text{out}\rangle$, given that it was originally in some initial state $|\text{in}\rangle$, using a “path integral” of $e^{iS/\hbar}$ over all possible histories that connect $|\text{in}\rangle$ and $|\text{out}\rangle$. The details are not important for us (and in any case are mathematically dubious in general). The intuition behind the path integral formulation is that the particle simultaneously follows every one of those histories, and each of them contributes to the probability. \hbar is called *Planck’s constant*; in the “classical limit” as $\hbar \rightarrow 0$, the contribution from the path that satisfies Hamilton’s principle dominates everything else.

The main use of Feynman’s path integral is in perturbation theory. Finding exact solutions in physics is typically impossible and rarely useful. In practice, it suffices to find the first few terms in some Taylor expansion of the solution. This so-called “perturbative” approach to quantum theories is particularly transparent in Feynman’s formalism, where each term of the expansion can be represented pictorially as a graph. See figure 1(a) for typical examples. The graphs contributing to the n th-order term in this Taylor expansion will involve n vertices. Feynman’s rules describe how to convert these graphs into integral expressions for computing the individual terms in the Taylor expansion.

In this article we are interested in perturbative string theory. The string Feynman diagrams (see figure 1(b) for three equivalent ones) are surfaces called *worldsheets*; the need for quantum foam is avoided because these surfaces are much less singular than the particle graphs (which have singularities at each vertex), and

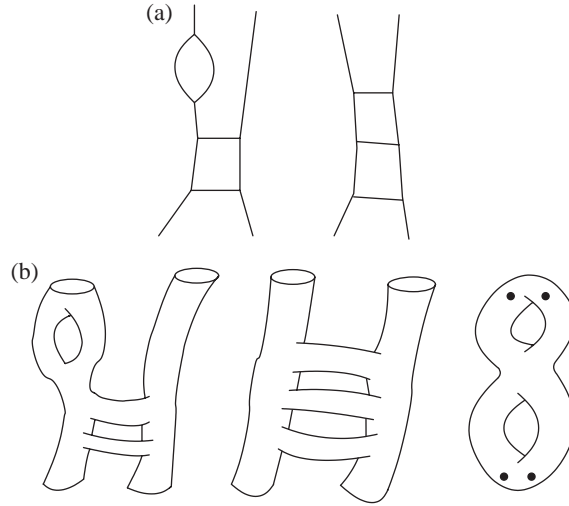


Figure 1 Some Feynman diagrams of (a) particles and (b) strings.

this is also largely why the mathematics of strings is so nice. To cut a long story short, each term in the perturbative expression for probabilities in string theory can be calculated from a quantity called a “correlation function” in a CFT that lives on the corresponding worldsheet. Feynman’s path integral here amounts to the integral of a quantity that CFT can compute, over some MODULI SPACE [IV.8] of surfaces.

The vertices in a Feynman diagram represent places where one particle absorbs or emits another. The corresponding rules of string theory tell us that we should dissect the worldsheet into “tubular Y-shapes,” or spheres with three legs, as in figure 2. Since these spheres with legs play the role of vertices in the Feynman diagram, the factor they contribute to the integrand of the path integral is called a *vertex operator*, and now it describes the absorption or emission of one *string* by another. A vertex operator algebra is the “algebra” of these vertex operators.

2.1.3 The Hamiltonian Formulation and Algebra

The Poisson bracket $\{A, B\}_P$ of two classical observables A and B is defined to be

$$\frac{\partial A}{\partial x} \frac{\partial B}{\partial p} - \frac{\partial B}{\partial x} \frac{\partial A}{\partial p}.$$

Note that $\{A, B\}_P = -\{B, A\}_P$: in other words, the Poisson bracket is *anti-commutative*. It also satisfies the *Jacobi identity*

$$\{A, \{B, C\}_P\}_P + \{B, \{C, A\}_P\}_P + \{C, \{A, B\}_P\}_P = 0,$$

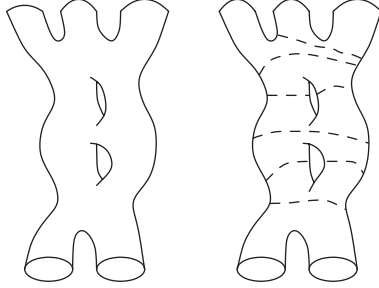


Figure 2 Dissecting a surface.

and therefore defines a Lie algebra. The Hamiltonian formulation of classical physics expresses the evolution of an observable A by means of the differential equation $\dot{A} = \{A, H\}_P$, where H is the HAMILTONIAN [III.35]: that is, the energy observable. The quantum version of this picture is due to Heisenberg and Dirac: the observables are now linear operators rather than smooth functions, and the Poisson bracket is replaced by the *commutator* $[\hat{A}, \hat{B}] = \hat{A} \circ \hat{B} - \hat{B} \circ \hat{A}$ of operators. This again has the anti-commuting property $[\hat{A}, \hat{B}] = -[\hat{B}, \hat{A}]$ and again satisfies the Jacobi identity, so the process of “quantization” gives rise to a homomorphism of Lie algebras. The derivative with respect to time of a quantum observable \hat{A} is then the natural analogue of the classical case: it is proportional to $[\hat{A}, \hat{H}]$, where \hat{H} is the Hamiltonian operator. Thus the Hamiltonian has a dual role: as the energy observable and as the controller of time evolution. All of physics is stored in the action of the observables on state space \mathcal{H} , as well as the commutators of these observables with \hat{H} .

Let us illustrate this picture with the *quantum spring*, also known as the *harmonic oscillator*. The position and momentum observables \hat{x} , \hat{p} are operators acting on the infinite-dimensional space \mathcal{H} of possible spring-states. It is more convenient to work with certain combinations of them called \hat{a} and \hat{a}^\dagger (the dagger denotes the “Hermitian adjoint,” or complex-conjugate transpose), which obey the commutator relation $[\hat{a}, \hat{a}^\dagger] = I$, where I is the identity operator. It turns out that all other observables can be built from \hat{a} and \hat{a}^\dagger . For example, the Hamiltonian \hat{H} is $l(\hat{a}^\dagger \hat{a} + \frac{1}{2})$ for some positive constant l . The *vacuum*, which is denoted $|0\rangle$, is the state of minimum energy. In other words, the state $|0\rangle$ is an eigenvector of \hat{H} with smallest possible eigenvalue: $\hat{H}|0\rangle = E_0|0\rangle$ for some $E_0 \in \mathbb{R}$ and all other eigenvalue E of \hat{H} are greater than E_0 . It follows from this that

$$\begin{aligned} \hat{a}|0\rangle &= 0. \text{ To see why, consider the effect of } \hat{H} \text{ on } \hat{a}|0\rangle: \\ \hat{H}\hat{a}|0\rangle &= l(\hat{a}^\dagger \hat{a} + \tfrac{1}{2})\hat{a}|0\rangle = l(\hat{a}\hat{a}^\dagger - \tfrac{1}{2})\hat{a}|0\rangle \\ &= \hat{a}l(\hat{a}^\dagger \hat{a} - \tfrac{1}{2})|0\rangle = \hat{a}(\hat{H} - l)|0\rangle = (E_0 - l)\hat{a}|0\rangle. \end{aligned}$$

Here, we have used the fact that $\hat{a}^\dagger \hat{a} = \hat{a}\hat{a}^\dagger - I$. (The observables \hat{a} and \hat{a}^\dagger are called *creation and annihilation operators* because, as we shall see later, they can be interpreted as adding or removing a particle from a certain n -particle state. Showing this uses the fact that they produce $\pm I$ when you interchange their order.) This calculation shows that if $\hat{a}|0\rangle$ is not zero, then it is an eigenvector of \hat{H} with an eigenvalue smaller than E_0 , which is a contradiction.

Since $\hat{a}|0\rangle = 0$, it follows that $\hat{H}|0\rangle = \frac{1}{2}l|0\rangle$, so $E_0 = \frac{1}{2}l$. We now define, for each positive integer n , a state $|n\rangle$ to be $(\hat{a}^\dagger)^n|0\rangle \in \mathcal{H}$. Similar calculations to the one just given show that $|n\rangle$ has energy $E_n = (2n + 1)E_0$. For example,

$$\begin{aligned} \hat{H}|1\rangle &= l(\hat{a}^\dagger \hat{a} + \tfrac{1}{2})\hat{a}^\dagger|0\rangle = l(\hat{a}^\dagger(\hat{a}^\dagger \hat{a} + I) + \tfrac{1}{2}\hat{a}^\dagger)|0\rangle \\ &= \tfrac{3}{2}l\hat{a}^\dagger|0\rangle = E_1|1\rangle. \end{aligned}$$

(Note that we used the fact that $\hat{a}|0\rangle = 0$ in the penultimate equality above.) We think of the vacuum as the ground state, and $|n\rangle$ as being the state with n *quantum particles*. These states $|n\rangle$ span all of the state space \mathcal{H} . To see how some observable acts on some state, one writes the observable in terms of the basic observables \hat{a} , \hat{a}^\dagger and the state in terms of the basic states $|n\rangle$. In this algebraic way we can recover all of the physics.

This idea of building up the whole space \mathcal{H} from the vacuum and the operators is a fruitful one in mathematics as well: something similar happens for the most important modules of most of the important Lie algebras.

2.1.4 Fields

A classical *field* is a function of space and time. Its values can be numbers or vectors, which represent quantities such as air temperature or the current in a river. The values taken by a *quantum field* are operators; furthermore, a quantum field is not a *function* of space and time, but a more general object called a *DISTRIBUTION* [III.18]. The prototypical example of a distribution is the *Dirac delta function* $\delta(x - a)$. Despite its name, this is not a function: rather, it is defined by the property that

$$\int f(x) \delta(x - a) dx = f(a) \quad (1)$$

for any sufficiently well-behaved function $f(x)$. Even though $\delta(x - a)$ is not a function, one can informally

interpret it as the derivative of a step function, and one can visualize it as equaling 0 everywhere except at $x = a$, where it is infinite, in such a way that the infinitely tall and infinitely thin rectangle under the graph has area 1. However, it really only makes sense inside an integral, as in (1). Similar remarks apply to distributions in general, so a quantum field can really only be evaluated inside an integral of space and time, applied to some “test function” like f above. The value of such an integral will be an operator on the state space \mathcal{H} .

Dirac deltas appear in classical mechanics when one takes Poisson brackets of classical fields. Similarly, commutators of quantum fields involve delta functions too. For example, in the simplest cases the quantum fields φ satisfy

$$\left. \begin{aligned} \varphi(x, t), \varphi(x', t) &= 0, \\ \left[\varphi(x, t), \frac{\partial}{\partial t} \varphi(x', t) \right] &= i\hbar \delta(x - x'). \end{aligned} \right\} \quad (2)$$

This is a mathematical way of expressing, in the context of quantum field theory, the cherished physical principle called *locality*:¹ the only way we can *directly* affect something is by nudging it. In order to influence something not touching us, we must propagate a disturbance from us to it, such as a ripple in water. The main purpose of both classical and quantum fields is that they provide a natural vehicle for realizing locality. Locality is also at the heart of vertex operator algebras.

An important aspect of modern physics is that many of the central concepts of classical physics become less central, and are instead *derived* quantities. For example, the basic object of GENERAL RELATIVITY [IV.13] is a Lorentzian manifold, and familiar physical quantities such as mass and gravitational force are, from the point of view of this manifold, just names (that are not wholly precise) given to certain of its geometrical features.

Particles are obviously essential to classical physics, but we have not mentioned them in our brief sketch of quantum field theory. They arise through the so-called *modes* of quantum fields φ , which play the role of the operators \hat{a}, \hat{a}^\dagger that we met in section 2.1.3. A mode is the operator that results from hitting the quantum field with an appropriate test function and integrating—just as one does when working out a Fourier coefficient,

in which case the test functions are TRIGONOMETRIC FUNCTIONS [III.94]. In fact, when viewed appropriately, modes actually *are* Fourier coefficients of a certain kind. The commutators of these modes can be obtained from the commutators of the fields. Now, recall that the vertex operators of string theory are related to the emission and absorption of strings. As we shall see shortly, these vertex operators are the quantum fields in a quantum field theory of point particles (namely, the associated conformal field theory); the modes of these vertex operators generate the “particles” (or in more conventional language, the *states*) in that conformal field theory. Equivalently, they generate the various vibrational states of a single string in that string theory.

2.2 Conformal Field Theory

A *conformal field theory* (CFT) is a quantum field theory with a two-dimensional space-time whose symmetries include all *conformal transformations*. We shall explain what this means in the next paragraphs, but for now it is enough to know that a CFT is a particularly symmetrical kind of quantum field theory. A CFT lives on the worldsheet Σ traced by a set of strings as they evolve, sometimes colliding and separating, through time. In this subsection we shall informally sketch their basic theory; in section 3.1 we shall be more precise.

CFT, like any quantum field theory in two dimensions, has two almost independent halves. This is easiest to see in the context of string theory: the ripples on the string are responsible for the physical properties (charge, mass, etc.) of the corresponding state, but they can move (at the speed of light) either clockwise or counterclockwise around the string. When they do so, they just pass through each other without interacting. These two alternatives, clockwise and counterclockwise, yield the two *chiral halves* of CFT. To study a CFT, one first analyzes its chiral halves and then splices them together to form the “bichiral” physical quantities. Almost all attention in CFT by mathematicians has focused on the chiral (as opposed to physical) data, and indeed that is where vertex operator algebras live. For ease of presentation, we will usually suppress one of the chiral halves.

A conformal transformation is a transformation that preserves angles. The simplest reason one can give for why two dimensions are so special for CFT is that there are far more conformal transformations in two dimensions than there are in higher dimensions. When $n > 2$ the only examples are the obvious ones: combinations of translations, rotations, and enlargements.

1. More precisely, for quantum fields, locality takes the form that if not even light can connect two given space-time points, then the quantum fields at those points must be causally independent. In particular, measurements at such points can be performed simultaneously with arbitrary precision. In quantum theories, this requires those operators to commute. Equation (2) is a generous way to satisfy locality.

PUP: brace on right-hand side is a way that we commonly make clear that two lines of a display are both associated with the same equation number. OK?

This means that the space of all local conformal transformations in \mathbb{R}^n is $\binom{n+2}{2}$ dimensional. However, when $n = 2$ the space of local conformal transformations is far richer: it is *infinite* dimensional. Indeed, if you identify \mathbb{R}^2 with the complex plane \mathbb{C} , then any HOLOMORPHIC FUNCTION [I.3 §5.6] $f(z)$ that does not have zero derivative at a point z_0 is conformal near z_0 . Since a CFT is invariant under conformal transformations and there are many conformal transformations, a CFT is especially symmetrical: this is what makes CFTs so interesting mathematically.

Lie algebras arise naturally whenever one has local symmetries, and indeed one can form an infinite-dimensional Lie algebra out of the infinitesimal conformal transformations. This algebra has a basis l_n , $n \in \mathbb{Z}$, that obeys the Lie-bracket relations

$$[l_m, l_n] = (m - n)l_{m+n}. \quad (3)$$

The algebraic interpretation of the conformal symmetry of CFT turns out to be that these basis elements l_n act naturally on all the quantities in the theory, as we shall explain below.

The basic example that underlies all the others is when space-time Σ is a semi-infinite cylinder corresponding to an incoming string. It is parametrized by time $t < 0$ and the angle $0 \leq \theta < 2\pi$ around the string. We can conformally map the cylinder to the punctured disk in \mathbb{C} by $z = e^{t-i\theta}$, so $t = -\infty$ corresponds to $z = 0$. This allows us to say what we mean by conformal symmetries of the cylinder.

The quantum fields $\varphi(z)$ of CFT are the vertex operators of string theory. As always, these quantum fields φ are “operator-valued distributions” on space-time Σ , acting on the space \mathcal{H} of states. Now it is possible for a field φ to be “holomorphic,” in the following sense. First, you calculate its modes φ_n , one for each $n \in \mathbb{Z}$, which are linear maps from the state space \mathcal{H} to itself, given by the formula

$$\varphi_n = \int \varphi(z) z^{n-1} dz,$$

where the integral is around a small circle about the origin. Then you take these modes as the coefficients of a formal power series $\sum_{n \in \mathbb{Z}} \varphi_n z^n$. We call φ holomorphic if this formal power series can be identified with φ , in a sense that we shall discuss more in section 3.1. A typical field $\varphi(z)$ is not holomorphic: rather, it is a combination of holomorphic and anti-holomorphic fields, which make up the two chiral halves of CFT. We will focus on the space of holomorphic fields $\varphi(z)$, which

we call \mathcal{V} . This turns out to form a vertex operator algebra (as do the anti-holomorphic fields).

For example, the most important vertex operator comes directly from the conformal symmetry: the *stress-energy tensor* $T(z) \in \mathcal{V}$ is the “conserved current” that Noether’s theorem associates with the conformal symmetry. Labeling its modes (Noether’s “conserved charges” here) by $L_n = \int T(z) z^{-n-3} dz$, so that $T(z) = \sum_n L_n z^{-n-2}$, we find that they *almost* realize the conformal algebra: instead of (3), however, they obey the slightly more complicated relations

$$[L_m, L_n] = (m - n)L_{m+n} + \delta_{n,-m} \frac{m(m^2 - 1)}{12} cI, \quad (4)$$

where I is the identity. In other words, the operators L_n and I form an extension of the conformal algebra by I . The resulting infinite-dimensional Lie algebra is called the *Virasoro algebra* \mathfrak{Vir} . The number c appearing in (4) is called the *central charge* of the CFT and is a rough measure of its size.

The operators L_n do not precisely represent the conformal algebra (3). Instead, they form a so-called *projective representation*. Projective representations of symmetries, such as (4), are common in quantum theories. The fact that they are not true representations is not a problem, since one can turn them into true representations by extending the algebra. In our case, the state space \mathcal{H} carries inside it a true representation of the Virasoro algebra \mathfrak{Vir} , which is useful as it means \mathfrak{Vir} can be used to organize \mathcal{H} .

Any quantum field theory has what is called a *state-field correspondence*: with each field φ one associates its incoming state, which is the limit as the time t tends to $-\infty$ of $\varphi|0\rangle$ (as always, $|0\rangle$ is the vacuum state in \mathcal{H} and φ acts on states). CFT is unusual in that the state-field correspondence is a bijection. This means we can identify \mathcal{H} and \mathcal{V} and use states to label all fields.

We want to make \mathcal{V} into some sort of algebra, but the obvious direct approach of taking products $\varphi_1(z)\varphi_2(z)$ fails, since distributions, unlike true functions, cannot in general be multiplied. For example, the Dirac delta $\delta(x - a)$ cannot be squared without causing problems in (1). However, even if the product $\varphi_1(z)\varphi_2(z)$ does not make sense, one can make sense of $\varphi_1(z_1)\varphi_2(z_2)$ as an operator-valued distribution on Σ^2 . It is then possible to recover most of the physics of CFT by studying the singular terms as $z_2 \rightarrow z_1$. By the *operator product expansion*, we mean expanding products $\varphi_1(z_1)\varphi_2(z_2)$ as sums of the form $\sum_h (z_1 - z_2)^h O_h(z_1)$. The set \mathcal{V} is closed under this

product in the sense that each coefficient $O_h(z)$ lies in \mathcal{V} . A typical example is

$$T(z_1)T(z_2) = \frac{1}{2}c(z_1 - z_2)^{-4}I + 2(z_1 - z_2)^{-2}T(z_1) \\ + (z_1 - z_2)\frac{d}{dz}T(z_1) + \cdots$$

Physicists call \mathcal{V} a *chiral algebra*; for us it is the prototypical example of a vertex operator algebra. It is not an algebra in the conventional sense though, since, given vertex operators $\varphi_1(z)$ and $\varphi_2(z)$, we have not just a single product $\varphi_1(z) * \varphi_2(z)$ in \mathcal{V} but infinitely many products $\varphi_1(z) *_h \varphi_2(z) = O_h(z)$, all belonging to \mathcal{V} .

The Hamiltonian plays a crucial role in any quantum field theory; here it turns out to be proportional to the mode L_0 discussed earlier. Being an observable, L_0 is diagonalizable on \mathcal{H} , which means that any state $v \in \mathcal{H}$ can be written as a sum $\sum_h v_h$, where $v_h \in \mathcal{H}$ has energy h : that is, $L_0 v_h = h v_h$.

There is a special class of CFT that is particularly well-behaved. Let $\tilde{\mathcal{V}}$ denote the space of all anti-holomorphic fields in the CFT—it is the other chiral half. Recall that the full CFT consists of \mathcal{V} and $\tilde{\mathcal{V}}$ spliced together. We call the CFT *rational* if $\mathcal{V} \oplus \tilde{\mathcal{V}}$ is so large that it has finite index, in an appropriate sense, in the full space of quantum fields in the CFT. The name “rational” arises because the central charge c and other parameters in a rational CFT have to be rational numbers.

The mathematics of rational CFT is especially rich. Let us briefly look at one example. (We will use several words that will be unfamiliar to most readers, but at least it will give some idea of which areas are touched by CFT.) As with everything else, the quantum probabilities arising in CFT are found by first computing chiral quantities and splicing them together. These chiral quantities are called *conformal* or *chiral blocks*, and are found using simple Feynman-like rules applied to dissections like figure 2. In rational CFT we get a finite-dimensional space $\mathcal{F}_{g,n}$ of chiral blocks for any world-sheet Σ , i.e., for any choice of genus g and number n of punctures. These spaces carry projective representations of the mapping class group $\Gamma_{g,n}$ (defined to be the fundamental group π_1 of the moduli space $\mathcal{M}_{g,n}$). This $\Gamma_{g,n}$ -representation is the source, for instance, of Jones’s relation of the BRAID GROUP [III.4] (and hence KNOTS [III.46]) to subfactors, Borchers’s explanation of “Monstrous Moonshine,” the Drinfel’d-Kohno monodromy theorem, and the modularity of affine Kac-Moody characters. Some of this we will touch on in section 4.

The most important example here is the torus, where the chiral blocks are *modular functions*, a class of functions of fundamental mathematical importance. A modular function is a meromorphic function (that is, a function that is holomorphic except at a few “poles” where it can tend to infinity) $f(\tau)$ that is defined on the upper half-plane $\mathbb{H} = \{\tau \in \mathbb{C} \mid \text{Im } \tau > 0\}$ and that is “symmetric” with respect to the group $\text{SL}_2(\mathbb{Z})$ of 2×2 matrices with integer entries and determinant 1, in the sense that for any such matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ the function $f(\tau)$ is closely related (though not necessarily exactly equal) to the function $f((a\tau + b)/(c\tau + d))$. We shall discuss this further in section 3.2.

The appearance of modularity can be understood by recalling from section 2.1.2 that Feynman’s path integral in string theory is an integral over moduli spaces. The moduli space $\mathcal{M}_{1,0}$ for the torus can be written as the quotient of the half-plane \mathbb{H} by the action of $\text{SL}_2(\mathbb{Z})$. Therefore, if one lifts the integrand of Feynman’s integral from $\mathcal{M}_{1,0}$ to \mathbb{H} , one obtains a function $Z(\tau)$ that is invariant under $\text{SL}_2(\mathbb{Z})$ and hence modular. This integrand $Z(\tau)$ is a quadratic combination of the chiral blocks for the torus.

3 What VOAs Are

It is possible to give a fully axiomatic definition of vertex operator algebras. However, when one first encounters this definition (and not just the first time either) it can seem very complicated and arbitrary, and one is given no feel for the importance of VOAs. Our treatment below will be much more informal: this will clarify their importance even if it hides much of their complexity. Thanks to the previous section, it is possible to give a quick justification for VOAs: if you concede that CFT (or equivalently, perturbative string theory) is important, and if you have seen how closely related CFT is to VOAs, then you must concede that VOAs are important. However, this is not the whole story, as we shall see.

3.1 Their Definition

Let us begin by defining them in terms of other concepts that must themselves be defined: a vertex operator algebra is an algebra of vertex operators, or in other words the chiral algebra \mathcal{V} of a conformal field theory.

The most important thing to understand in this definition is that a vertex operator is a quantum field, which, as we have seen, is an “operator-valued distribution of space-time.” So we can think of it informally as a

PUP: I can confirm that this sentence is OK as written.

matrix-valued function of space-time, where the matrix is $\infty \times \infty$ and its entries can be generalized functions like the Dirac delta (1). However, we shall give a much better description of these vertex operators shortly.

By “space-time” we mean the unit disk in \mathbb{C} punctured at $z = 0$. Recall from section 2.2 that string-theoretically this set corresponds to a semi-infinite cylinder parametrized by the angle $-\pi < \theta \leq \pi$ running around the string as well as the time $-\infty < t < 0$ running along the axis: the map from this to the punctured disk was $(\theta, t) \mapsto z = e^{t-i\theta}$. We want to restrict our attention to quantum fields that depend holomorphically on z . However, it is not obvious what “holomorphic” means for distributions. We touched on this question in section 2.2: now we shall look at it in more detail.

To do this, we need a more concrete description of a vertex operator. The key idea is a very convenient algebraic interpretation of holomorphic distributions. Consider the sum

$$d(z) = \sum_{n=-\infty}^{\infty} z^n. \quad (5)$$

Multiply it by $f(z) = 3z^{-2} - 5z^3$, say. This gives us

$$\begin{aligned} f(z)d(z) &= 3 \sum_{n=-\infty}^{\infty} z^{n-2} - 5 \sum_{n=-\infty}^{\infty} z^{n+3} \\ &= 3 \sum_{n=-\infty}^{\infty} z^n - 5 \sum_{n=-\infty}^{\infty} z^n = -2d(z). \end{aligned}$$

A few more examples like this will convince you that $f(z)d(z) = f(1)d(z)$ for *any* polynomial function f of z and z^{-1} . Therefore, $d(z)$ behaves exactly like the Dirac delta $\delta(z-1)$, at least for polynomial test functions f . Note that $d(z)$ cannot converge for any z : the positive powers have a convergent sum only for $|z| < 1$, and the negative powers only for $|z| > 1$. The “function” $d(z)$ is an example of a *formal power series*: any series $\sum_{n=-\infty}^{\infty} a_n z^n$, where the coefficients a_n can be anything and we ignore all convergence issues.

By inspection, these formal power series are “holomorphic” throughout the punctured plane: after all, holomorphic just means that the complex derivative d/dz exists, and the derivative $\sum_n n a_n z^{n-1}$ of a formal power series clearly remains a formal power series. (By contrast, nonholomorphic series would involve the complex conjugate \bar{z} .)

So that is what a vertex operator looks like: a formal power series $\sum_{n=-\infty}^{\infty} a_n z^n$, where each coefficient a_n is now an operator (endomorphism) on the space \mathcal{V} of states, which is an infinite-dimensional vector space.

Since the vertex operators are in one-to-one correspondence with the states (we called this the “state-field correspondence” above), we can label these vertex operators with states: the standard convention is to denote the vertex operator corresponding to state $v \in \mathcal{V}$ by

$$Y(v, z) = \sum_{n=-\infty}^{\infty} v_n z^{-n-1}. \quad (6)$$

The symbol “ Y ” should remind you of the sphere with three legs, which as we know is the vertex of string theory. These coefficients v_n are the modes: as in any quantum field theory, all observables and states in the theory are built up from them.

The most important state in the theory is the vacuum $|0\rangle$. It corresponds to the identity vertex operator: $Y(|0\rangle, z) = I$. From the physical point of view, the vertex operator $Y(v, z)$ is the field that created the state v at time $t = -\infty$, i.e., $Y(v, 0)|0\rangle$ exists and equals v . (Recall that in our model $z = 0$ corresponds to $t = -\infty$.) Among other things, this means that $v_{-1}(|0\rangle) = v$, so indeed the modes applied to $|0\rangle$ generate \mathcal{V} , as is required in any quantum field theory.

The most important observable in the theory is the Hamiltonian, or energy operator, which we denote by L_0 . It is diagonalizable (so \mathcal{V} can be written as a sum of L_0 -eigenspaces) and all of its eigenvalues must be integers. For example, the vacuum $|0\rangle$ has 0 energy: $L_0|0\rangle = 0$. Since $|0\rangle$ should have the minimum energy, the L_0 -decomposition of \mathcal{V} is then $\mathcal{V} = \bigoplus_{n=0}^{\infty} \mathcal{V}_n$, where $\mathcal{V}_0 = \mathbb{C}|0\rangle$. Each space \mathcal{V}_n turns out to be finite dimensional, and we can think of L_0 as defining a \mathbb{Z}_+ -grading on state space \mathcal{V} .

The most important vertex operator in the theory is the stress-energy tensor $T(z)$. The corresponding state is called the *conformal vector* ω : $Y(\omega, z) = T(z)$. This means that ω has modes $\omega_n = L_{n-1}$ that form a representation (4) of the Virasoro algebra \mathfrak{Vir} . (This is the algebraic expression for the requirement of conformal symmetry.) The conformal vector has energy 2: $\omega \in \mathcal{V}_2$.

So far our theory is seriously underdetermined. The most important axiom to help us to pin it down further is locality. With a little work, one can show that this reduces to the condition that the commutator $[Y(u, z), Y(v, w)]$ of two vertex operators should be a finite linear combination of the Dirac delta $\delta(z-w) = z^{-1} \sum_{n=-\infty}^{\infty} (w/z)^n$ and its derivatives $(\partial^k/\partial w^k) \delta(z-w)$. Now, $(z-w)^{k+1} (\partial^k/\partial w^k) \delta(z-w) = 0$. To see

this, look at the case $k = 1$:

$$\begin{aligned} & (z-w)^2 \frac{\partial}{\partial w} \delta(z-w) \\ &= \sum_{n=-\infty}^{\infty} (nw^{n-1}z^{-n+1} - 2nw^n z^{-n} + nw^{n+1}z^{-n-1}) \\ &= \sum_{n=-\infty}^{\infty} ((n+1) - 2n + (n-1))w^n z^{-n} = 0. \end{aligned}$$

The proof for general k is similar. Therefore, locality can be recast in an equivalent form as follows: given any $u, v \in \mathcal{V}$, there is a positive number N such that

$$(z-w)^N [Y(u, z), Y(v, w)] = 0. \quad (7)$$

This equation may look strange. Why can we not simply divide out the $(z-w)^N$ and get that all vertex operators commute? The reason is that when formal power series are involved, there can be zero divisors. For example, it is easy to check that $(z-1) \sum_{n \in \mathbb{Z}} z^n = 0$. Locality in the form (7) is at the heart of VOAs; for instance, one can express it as a triply infinite sequence of identities that the modes must obey, and this emphasizes just how restrictive a condition it is, and how correspondingly interesting it is to find examples of VOAs.

This completes the definition of a VOA. A consequence of these properties is that the modes u_n respect the L_0 -grading that we mentioned earlier. This means that if u has energy k and v has energy l , then $u_n(v)$ has energy $k + l - n - 1$. The definition followed here is sometimes called a VOA of *CFT-type*, for obvious reasons. Sometimes in the literature some of these conditions are weakened or dropped. For example, much of the theory is independent of the existence of the conformal vector ω , although to us it will be crucial, for reasons that will be explained in the next subsection.

A VOA is simultaneously a physical and a mathematical object. We have emphasized their physical origins in order to help explain the motivation for studying them. We know they should be valuable to mathematicians, simply because CFT is, and indeed this is the case, as we shall see in section 4. But from a purely mathematical point of view, they might appear somewhat ad hoc, as though we had a list of mathematical ingredients and said to ourselves, “Let’s consider this, and then have some of these, oh, and perhaps one of those too, but with the following extra assumption: . . .” Fortunately, there are more abstract formulations of VOAs that make them appear much less arbitrary as mathematical structures. For example, Huang has shown that they can be regarded as “two-dimensionalized” Lie algebras, in the following sense. If you want to keep track of

the Lie brackets in an expression such as $[a, [[b, c], d]]$ (which is important since the Lie bracket is not an associative operation), you can do so with the help of a binary tree, and in fact it is easy to formulate Lie algebras in the language of such trees. If one then replaces binary trees by diagrams made out of spheres with legs, as we did with Feynman diagrams earlier, one obtains a structure that is equivalent to a VOA. (Of course, this is very far from a full explanation of what Huang did: his proof is extremely long.)

3.2 Basic Properties

We see from the definition sketched in the last subsection that a VOA is an infinite-dimensional \mathbb{Z}_+ -graded vector space with infinitely many products (namely $u * v = u_n(v)$), which obey infinitely many identities. Needless to say, it is not an easy definition, and there are no easy examples.

However, if we ignore the conformal symmetry (i.e., the conformal vector ω), then there are some simple, though uninteresting, examples. The easiest is the one-dimensional algebra $\mathcal{V} = \mathbb{C}|0\rangle$. More generally, a VOA \mathcal{V} that obeys (7) with $N = 0$ is a commutative associative algebra with a unit $1 = |0\rangle$. It also has a *derivation* $T = L_{-1}$, with respect to the product $u * v = u_{-1}(v)$: this means a linear map that obeys the product rule satisfied by derivatives, namely $T(u * v) = (Tu) * v + u * (Tv)$. The converse of this statement is true too: any such algebra is a VOA that obeys (7) with $N = 0$. In these simple examples, the role of the derivation T is to recover the z -dependence of the vertex operator.

Therefore, we need N not to be zero in (7) if we want interesting examples. Likewise, the vertex operators $Y(u, z)$ must be distributions (that is, they must involve doubly infinite sums) or again the VOA reduces to a commutative associative algebra.

It is also easy to show that in any VOA (again the existence of the conformal vector is not needed), the space \mathcal{V}_1 is a Lie algebra, with Lie bracket given by $[uv] = u_0(v)$. This is important because each \mathcal{V}_n will carry a representation of this Lie algebra, and \mathcal{V}_1 generates continuous symmetries of the VOA (at least when $\mathcal{V}_1 \neq \{0\}$). For a typical VOA \mathcal{V} these Lie algebras are very familiar. For instance, for the VOAs associated with rational CFT, they are *reductive*, which means that they are a direct sum of copies of the trivial Lie algebra \mathbb{C} with simple Lie algebras.

The existence of the conformal vector becomes important when one starts to consider the represen-

tation theory of VOAs. A \mathcal{V} -module is defined in a natural way. We shall not give full details here, but, roughly speaking, it is a space on which \mathcal{V} acts in such a way that as much as possible of the VOA structure is respected. For example, \mathcal{V} will automatically be a module for itself, just as a group acts on itself in a simple way. (See REPRESENTATION THEORY [IV.9 §2] for an explanation of the latter.) A *rational* VOA is defined to be one that has the simplest representation theory: it has only finitely many irreducible \mathcal{V} -modules, and any \mathcal{V} -module is a direct sum of irreducible ones. They are called rational VOAs because they are the VOAs that come from rational CFT. For these VOAs, \mathcal{V} acts irreducibly on itself.

Any irreducible \mathcal{V} -module M will inherit from \mathcal{V} an L_0 -grading by rational numbers, $M = \bigoplus_h M_h$, into finite-dimensional spaces M_h . The *character* $\chi_M(\tau)$ is defined by

$$\chi_M(\tau) = \sum_h \dim M_h e^{2\pi i \tau (h - c/24)}, \quad (8)$$

where c is the central charge. This definition arises naturally in CFT as well as in Lie theory (or *affine Kac-Moody algebras*), although the curious “ $c/24$,” needed for (9) below, is mysterious in Lie theory. (In CFT it has a natural explanation as a certain topological effect.) These characters converge for any τ in the upper half-plane \mathbb{H} . They carry a representation of the modular group $SL_2(\mathbb{Z})$:

$$\chi_M\left(\frac{a\tau + b}{c\tau + d}\right) = \sum_{N \in \Phi(\mathcal{V})} \rho\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \chi_N(\tau), \quad (9)$$

where, writing n for the number of irreducible \mathcal{V} -modules, the matrices $\rho\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ are $n \times n$ matrices with complex entries. The lengthy proof of (9), by Zhu, is perhaps the high point of VOA theory, and owes much to the intuitions of rational CFT. In the next section, we shall get some idea of why it is so important.

4 What Are VOAs Good For?

This section describes what are perhaps the two most significant applications of VOAs. But let us begin by listing (without any explanations) a few others. Inspired by the geometry of string theory, vertex operator (super)algebras have been assigned to manifolds, resulting in a powerful, though complicated, algebraic invariant of those manifolds that generalizes and enriches more classical data such as de Rham cohomology. VOAs associated with affine Kac-Moody algebras

at “degenerate” levels k are deeply related to the geometric Langlands program. The modularity of affine algebra characters, as well as that of, for example, lattice theta functions, are all special cases of Zhu’s theorem, which places these modularities in a much broader context.

4.1 The Mathematical Formulation of CFT

Starting in the 1970s, quantum field theory has had considerable success, especially in geometry, by studying classical structures using infinite-dimensional methods; this is a theme in particular of Atiyah’s school. Conformal field theories are a class of exceptionally symmetric quantum field theories, and they are also among the simplest nontrivial quantum field theories known. In the past two decades mathematics has feasted on this combination of symmetry and (relative) simplicity, often by “looping” or “complexifying” more classical structures, and the impact of CFT (or, equivalently, of string theory) has been especially significant and broad. In hindsight the importance of CFT to mathematics is not surprising: it is a coherent and intricate structure that straddles several disparate areas of mathematics, sprawling across geometry, number theory, analysis, combinatorics, and indeed algebra.

From this point of view, a crucial application of VOA theory has been to CFT itself. Quantum field theories are notoriously difficult to put on a rigorous mathematical footing. But the successful applications suggest that these difficulties are a symptom of mathematical profundity and subtlety rather than of irreparable mathematical incoherence. In this sense the situation is highly reminiscent of the deep conceptual challenges to eighteenth-century mathematicians that were raised by calculus. The definition of a VOA by Richard Borcherds makes the chiral algebra of a CFT completely rigorous, as well as concepts like the operator product expansion. Subsequent work (especially by Huang and Zhu) reconstructs from the VOA more and more of the CFT, in arbitrary genus. The resulting clarity makes the whole subject more accessible to, and hence exploitable by, mathematicians. Quantum field theories are here to stay in mathematics, and thanks to VOAs mathematicians are absorbing a large class of them completely and explicitly.

4.2 Monstrous Moonshine

In 1978 McKay noticed that $196884 \approx 196883$. Why was this an interesting observation? Well, the number

PUP: Tim thinks this reads OK as it makes the operator product expansion rigorous too. OK?

on the left is the first meaningful coefficient of the j -FUNCTION [IV.1 §8]

$$j(\tau) = q^{-1} + (744+)196\,884q + 21\,493\,760q^2 + 864\,299\,970q^3 + \cdots, \quad (10)$$

the generator of all modular functions for $\mathrm{SL}_2(\mathbb{Z})$. Recall that a modular function is a function $f(\tau)$ that is meromorphic in the upper half-plane \mathbb{H} and invariant under the usual action of $\mathrm{SL}_2(\mathbb{Z})$. It should also be meromorphic at the boundary points $\mathbb{Q} \cup \{\infty\}$, which are called *cusps*; we did not mention this condition earlier. The j -function generates these functions in the sense that any such modular function $f(\tau)$ can be written as a rational function $\mathrm{poly}(j(\tau))/\mathrm{poly}(j(\tau))$. In other words, $j(\tau)$ is a uniformizing function that identifies $(\mathbb{H} \cup \mathbb{Q} \cup \{\infty\})/\mathrm{SL}_2(\mathbb{Z})$ with the Riemann sphere $\mathbb{C} \cup \infty$. We bracketed the constant term 744 in (10) because although 744 was the traditional choice it can be freely replaced with any other number, including 0.

The number on the right in McKay's observation is the dimension of the smallest nontrivial representation of the Monster, the most exceptional of the FINITE SIMPLE GROUPS [V.8]. This relation between modular functions and the Monster was completely unexpected, as they seem to occupy completely independent spots in the mathematical universe. Conway, Norton, and others fleshed out and expanded McKay's original observation by making a number of conjectures, collectively called *Monstrous Moonshine*. For instance, with every pair (g, h) of commuting elements in the Monster (a group of size about 8×10^{53}), we expect there to be associated a function $j_{(g,h)}(\tau)$ that generates all modular functions for some discrete subgroup $\Gamma_{(g,h)}$ of $\mathrm{SL}_2(\mathbb{Z})$. The j -function would be assigned in the case $g = h = \text{identity}$.

The first major step toward proving these Moonshine conjectures was made by Frenkel, Lepowsky, and Meurman in the mid 1980s. They constructed an infinite-dimensional vector space V^\natural out of formal power series. They were motivated on the one hand by the vertex operators of string theory, and on the other by the formally similar distributions used in constructing affine algebra representations. This seemed a promising direction since for both string theory and affine algebra representations modular functions arise naturally. Together with a rich algebraic structure that came from these "vertex operators," V^\natural was also acted on in a natural way by the Monster group. Moreover, although V^\natural is infinite dimensional, it comes packaged into finite-dimensional pieces $V^\natural = \bigoplus_{n=-1}^{\infty} V_n^\natural$, and the

"graded dimension" $\sum_n \dim(V_n^\natural)q^n$ equals $j - 744$. The action of the Monster sends each V_n^\natural to itself; that is, each space V_n^\natural itself carries a representation of the Monster. Frenkel, Lepowsky, and Meurman proposed that V^\natural lies at the heart of the Monstrous Moonshine conjectures.

Borcherds was struck by the formal similarity between V^\natural and the chiral algebras of CFTs, and by abstracting out their important algebraic properties he defined a new structure called a vertex (operator) algebra. His axioms clarified their relationship with (generalizations of) Kac-Moody algebras, and by 1992 he had proved the main Conway-Norton conjecture (which corresponds to the case where g is arbitrary but h is the identity in the conjecture given earlier). Although his definition of VOAs required a deep understanding of the physics of CFT, his elaborate proof of this Moonshine conjecture is purely algebraic.

We would now call V^\natural a rational VOA with only one irreducible module (namely itself); its symmetry group is the Monster and its character (8) is $j(\tau) - 744$. The removal of the constant term 744 from (10) is significant as it says that the Lie algebra V_1^\natural is trivial—this is necessary if the symmetry group is to be finite. It is conjectured that V^\natural is the unique VOA with central charge $c = 24$, trivial \mathcal{V}_1 , and only one irreducible module. This is meant to be reminiscent of the LEECH LATTICE [I.4 §4], which is known to be the unique twenty-four-dimensional even self-dual lattice with no vectors of length $\sqrt{2}$. Indeed, the Leech lattice plays a crucial role in the construction of V^\natural .

Most of the Moonshine conjectures are still open and this deep connection between modular functions and the Monster is still somewhat mysterious. At the time of writing, however, VOAs still provide the only serious approach to the Moonshine conjectures.

Borcherds defined VOAs to clarify the chiral algebra of CFT and to tackle Monstrous Moonshine. For this work, he was awarded a Fields Medal in 1998.

Further Reading

- Borcherds, R. E. 1986. Vertex algebras, Kac-Moody algebras, and the Monster. *Proceedings of the National Academy of Sciences of the USA* 83:3068–71.
- . 1992. Monstrous Moonshine and monstrous Lie superalgebras. *Inventiones Mathematicae* 109:405–44.
- Di Francesco, P., P. Mathieu, and D. Sénéchal. 1996. *Conformal Field Theory*. New York: Springer.
- Gannon, T. 2006. *Moonshine Beyond the Monster: The Bridge Connecting Algebra, Modular Forms and Physics*. Cambridge: Cambridge University Press.

- Kac, V. G. 1998. *Vertex Algebras for Beginners*, 2nd edn. Providence, RI: American Mathematical Society.
- Lepowsky, J., and H. Li. 2004. *Introduction to Vertex Operator Algebras and their Representations*. Boston, MA: Birkhäuser.

IV.18 Enumerative and Algebraic Combinatorics

Doron Zeilberger

1 Introduction

Enumeration, otherwise known as *counting*, is the oldest mathematical subject, while algebraic combinatorics is one of the youngest. Some cynics claim that algebraic combinatorics is not really a new *subject* but just a new *name* given to enumerative combinatorics in order to enhance its (former) poor image, but algebraic combinatorics is in fact the synthesis of two opposing trends: *abstraction of the concrete* and *concretization of the abstract*. The former trend dominated the first half of the twentieth century, starting with Hilbert's "theological" proof of the fundamental theorem of invariants, in which he showed by abstract means that certain invariants existed, but not how to find them. The latter trend is dominating contemporary mathematics, thanks to the omnipresence of The Mighty Computer.

The abstraction trend consists of the *categorization*, *conceptualization*, *structuralization*, and *fancification* (in short, "BOURBAKIZATION" [VI.96]) of mathematics. Enumeration did not escape this trend, and in the hands of such giants as Gian-Carlo Rota and Richard Stanley in America and Marco Schützenberger and Dominique Foata in France, classical, enumerative combinatorics became more conceptual, structural, and algebraic. However, as algebraic combinatorics has established itself as a fully fledged and separate mathematical speciality, the more recent trend toward the *explicit*, *concrete*, and *constructive* has left its mark as well. It has revealed that many algebraic structures have hidden combinatorial underpinnings; the attempts to unearth these have led to many fascinating discoveries and unsolved problems.

1.1 Enumeration

The fundamental theorem of enumeration, independently discovered by several anonymous cave dwellers, states that

$$|A| = \sum_{a \in A} 1.$$

In words: the number of elements in A is the sum over all elements of A of the constant function 1.

While this formula is still useful after all these years, enumerating specific finite sets is no longer considered mathematics. A genuine mathematical fact has to incorporate *infinitely* many facts, and the generic enumeration problem is to enumerate not just one set but all the sets in an infinite family.

To be precise, given an infinite sequence of sets $\{A_n\}_{n=0}^{\infty}$, where each set A_n consists of objects satisfying some combinatorial specifications that depend on the parameter n , answer the question: How many elements does A_n have?

In a moment we shall look at some examples. But before we can learn how to *answer* this kind of question, let us consider a meta-question: What is an answer?

This was posed, and beautifully answered, by Herbert Wilf. To give some background to Wilf's meta-answer, let us examine answers to some famous instances of enumeration questions.

In the list below, when we are given a set A_n (which will change from example to example), we shall write a_n instead of $|A_n|$. That is, a_n will stand for the number of elements of A_n .

- (i) **I Ching**. If A_n is the set of all subsets of $\{1, \dots, n\}$, then $a_n = 2^n$.
- (ii) **Rabbi Levi Ben Gerson**. If A_n is the set of PERMUTATIONS [III.70] on $\{1, \dots, n\}$, then $a_n = n!$.
- (iii) **Catalan**. If A_n is the set of legal bracketings with n opening brackets and n closing brackets, then $a_n = (2n)!/(n+1)n!$. (A *legal bracketing* is a sequence of n opening brackets and n closing brackets such that at no point in the sequence has the number of closing brackets exceeded the number of opening brackets. For instance, when $n = 2$ the legal bracketings are $[[]]$ and $[[]]$.)
- (iv) **LEONARDO OF PISA [VI.6]**. Let A_n be the set of finite sequences that consist only of 1s and 2s and that sum to n . (For example, when $n = 4$ the possible sequences are 1111, 112, 121, 211, and 22.) In this case, we have *three* equivalent answers as follows.

(i)

$$a_n = \frac{1}{\sqrt{5}} \left(\left(\frac{1+\sqrt{5}}{2} \right)^{n+1} - \left(\frac{1-\sqrt{5}}{2} \right)^{n+1} \right).$$

(ii)

$$a_n = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k}.$$

Note to PUP:
proofreader
marked a correctly
matched
parenthesis for
deletion here. I
presume it's OK to
leave it as it is?

- (iii) $a_n = F_{n+1}$, where F_n is the sequence defined by the recurrence $F_n = F_{n-1} + F_{n-2}$, subject to the initial conditions $F_0 = 0, F_1 = 1$.
- (v) **CAYLEY [VI.46]**. If A_n is the set of labeled trees on n vertices, then $a_n = n^{n-2}$. (A *tree* is a connected GRAPH [III.34] without cycles, and it is *labeled* if the vertices have distinct names.)
- (vi) If A_n is the set of labeled simple graphs with n vertices, then $a_n = 2^{n(n-1)/2}$. (A graph is *simple* if it has neither loops nor multiple edges.)
- (vii) If A_n is the set of labeled *connected* simple graphs on n vertices (that is, graphs for which every vertex can be reached from every other by a path), then a_n is $n!$ times the coefficient of x^n in the power series expansion of

$$\log \left(\sum_{k=0}^{\infty} \frac{2^{k(k-1)/2}}{k!} x^k \right).$$

- (viii) If A_n is the number of Latin squares of size n ($n \times n$ matrices each of whose rows and columns is a permutation of $\{1, \dots, n\}$), then $a_n = \text{????}$.

In 1982, Wilf defined an answer as follows.

Definition. An *answer* is a *polynomial-time algorithm* (in n) for computing a_n .

Wilf arrived at this definition after he refereed a paper proposing a “formula” for the answer to question (viii), and realized that its “computational complexity” exceeds that of the caveman’s formula of direct counting.

What is a “formula”? It is really an algorithm that inputs n and outputs a_n . For example, $a_n = 2^n$ is shorthand for the recursive algorithm

$$\begin{aligned} &\text{if } n = 0 \text{ then } a_n = 1, \\ &\text{else } a_n = 2 \cdot a_{n-1}, \end{aligned}$$

which takes $O(n)$ steps. However, using the algorithm

$$\begin{aligned} &\text{if } n = 0 \text{ then } a_n = 1, \\ &\text{else if } n \text{ is odd, then } a_n = 2a_{n-1}, \\ &\text{else } a_n = a_{n/2}^2 \end{aligned}$$

takes $O(\log n)$ steps, much faster than Wilf demands. In other cases, like enumerating self-avoiding walks, the best algorithm that is known is exponential, $O(c^n)$, and any lowering of the constant c is a major advance. (A *self-avoiding walk* is a sequence of points $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_n$ in the two-dimensional integer lattice, where each \mathbf{x}_i is one of the four neighbors of \mathbf{x}_{i-1} and no two of the \mathbf{x}_i are equal.) Notwithstanding these

exceptions, Wilf’s meta-answer is a very useful general guideline for evaluating answers.

The traditional customers of enumeration were mainly probability and statistics. In fact, discrete probability is almost synonymous with enumerative combinatorics, since the probability of an event E occurring is the ratio of the number of successful cases divided by the total number. Also, statistical physics is, by and large, weighted enumeration of lattice models (see PHASE TRANSITIONS AND UNIVERSALITY [IV.25]). About fifty years ago, another important customer came along: computer science. Here one is interested in the COMPUTATIONAL COMPLEXITY [IV.20] of algorithms: that is, in the number of steps it takes to execute them.

2 Methods

The following tools are indispensable to the enumerative combinatorialist.

2.1 Decomposition

$$|A \cup B| = |A| + |B| \quad (\text{if } A \cap B = \emptyset).$$

In words: the size of the union of two disjoint sets equals the sum of their sizes.

$$|A \times B| = |A| \cdot |B|.$$

In words: the size of the Cartesian product of two sets (that is, the set of all pairs (a, b) , where $a \in A$ and $b \in B$) equals the product of their sizes.

$$|A^B| = |A|^{|B|}.$$

In words: the size of the set of functions from B to A equals the size of A raised to the power the size of B . For example, the number of 0-1 sequences of length n , which can be viewed as functions from $\{1, 2, \dots, n\}$ to $\{0, 1\}$, equals 2^n .

2.2 Refinement

If

$$A_n = \bigcup_k B_{nk} \quad (\text{disjoint union}),$$

and if b_{nk} , the number of elements of B_{nk} , is “nice” (and even if it is not), then

$$a_n = \sum_k b_{nk}.$$

The idea here is that it may be possible to take a set A_n that is difficult to count, and split it up into disjoint sets B_{nk} that are easier to count. For example, consider the

PUP: difficult to get across the uncertainty about this result but what do you think of the extra question marks here, to make it clearer that it's not just a mistake - we hope!

set A_n of example (iv). This can be split into a disjoint union of subsets B_{nk} , where each B_{nk} consists of the sequences in A_n that have exactly k 2s. If there are k 2s, then there must be $n - 2k$ 1s, so $b_{nk} = \binom{n-k}{k}$. This yields answer (ii).

2.3 Recursion

Suppose that A_n can be decomposed in such a way that it is a combination of fundamental operations applied to the sets $A_{n-1}, A_{n-2}, \dots, A_0$. Then a_n satisfies a recurrence relation of the form

$$a_n = P(a_{n-1}, a_{n-2}, \dots, a_0).$$

For example, let A_n be the set of example (iv). If a sequence in A_n starts with a 1, then the rest of the sequence must add up to $n - 1$, and if it starts with a 2, then the rest must add up to $n - 2$. Since when $n \geq 2$ exactly one of these possibilities occurs and both are possible, we can decompose A_n into $1A_{n-1}$ and $2A_{n-2}$, where $1A_{n-1}$ is shorthand for the set of all sequences that begin with a 1 and continue with a sequence in A_{n-1} , and $2A_{n-2}$ is defined similarly. Since the sizes of $1A_{n-1}$ and $2A_{n-2}$ are clearly a_{n-1} and a_{n-2} , it follows that $a_n = a_{n-1} + a_{n-2}$, which yields answer (iii).

If A_n is the set of legal bracketings with n pairs (example (iii)), then a typical legal bracketing can be written recursively as $[L_1]L_2$, where L_1 and L_2 are smaller (possibly empty) legal bracketings. For example, if the bracketing is $[[]][[]][[]][[]]$ then $L_1 = [[]]$ and $L_2 = [[][[]][[]]]$. If L_1 has k pairs, then L_2 has $n - 1 - k$ pairs. It follows that A_n can be identified with the union $\bigcup_{k=0}^{n-1} A_k \times A_{n-1-k}$, and, taking cardinalities, $a_n = \sum_{k=0}^{n-1} a_k a_{n-1-k}$. This is a *nonlinear* (in fact, quadratic) and *nonlocal* recurrence, but it is nevertheless one that satisfies Wilf's dictum.

2.4 Generatingfunctionology

According to Wilf, who coined this neologism by making it the title of his classic book (a free download from his Web site, even though it is still in print!):

A generating function is a clothesline on which we hang up a sequence of numbers for display.

The method of generating functions is one of the most useful tools of the trade of enumeration. The generating function of a sequence, sometimes called its *z-transform*, is a discrete analogue of the LAPLACE TRANSFORM [III.93], and indeed goes back to LAPLACE [VI.23] himself. If the sequence is $(a_n)_{n=0}^{\infty}$, then its generating function $f(x)$ is defined to be $\sum_{n=0}^{\infty} a_n x^n$. In other

words, the terms of the sequence are regarded as the coefficients of a power series in x .

Generating functions are so useful because information about the sequence (a_n) translates to information about $f(x)$ that is often easier to process, and after some manipulations one often gets additional information about $f(x)$ that can be translated back into information about the sequence. For example, if $a_0 = a_1 = 1$ and $a_n = a_{n-1} + a_{n-2}$ when $n \geq 2$, then we can do the following manipulations on $f(x)$:

$$\begin{aligned} f(x) &= \sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + \sum_{n=2}^{\infty} a_n x^n \\ &= 1 + x + \sum_{n=2}^{\infty} (a_{n-1} + a_{n-2}) x^n \\ &= 1 + x + \sum_{n=2}^{\infty} a_{n-1} x^n + \sum_{n=2}^{\infty} a_{n-2} x^n \\ &= 1 + x + x \sum_{n=2}^{\infty} a_{n-1} x^{n-1} + x^2 \sum_{n=2}^{\infty} a_{n-2} x^{n-2} \\ &= 1 + x + x(f(x) - 1) + x^2 f(x) \\ &= 1 + (x + x^2)f(x). \end{aligned}$$

It follows that

$$f(x) = \frac{1}{1 - x - x^2}.$$

If one performs a partial-fraction decomposition, and expands the two resulting terms in a Taylor series, then one can obtain answer (i) to example (iv).

3 Weight Enumeration

According to the modern approach, pioneered by Pólya, Tutte, and Schützenberger, generating functions are neither “generating,” nor are they functions. Rather, they are *formal power series* that are *weight enumerators* of combinatorial sets. (Usually, but not always, these sets are infinite: for a finite set the corresponding “power series” has only finitely many nonzero terms and is therefore a polynomial.)

A power series $\sum_{n=0}^{\infty} a_n x^n$ is called *formal* when one sheds its analytical connotation as a Taylor series of a function, and thereby obviates the need to worry about convergence. For example, the sum $\sum_{n=0}^{\infty} n! x^n$ is perfectly legal as a formal power series even though it converges only when $x = 0$.

As for weight enumerators, consider the following situation. Suppose that we want to study the age distribution of a finite population. One way of doing this is to ask 121 questions. For each i between 0 and 120, we

ask those whose age is i to raise their hand. Then we count each of these age-groups one by one, compiling a table of a_i ($0 \leq i \leq 120$), and finally computing the generating function

$$f(x) = \sum_{i=0}^{120} a_i x^i.$$

But if the size of the population is much less than 120, it is much more efficient, because fewer questions would be needed, to ask every person their age and then to declare the *weight* of a person of age i to be x^i . Then the generating function is the sum of these weights. That is,

$$f(x) = \sum_{\text{persons}} x^{\text{age}(\text{person})},$$

which is a natural extension of the caveman's formula of naive counting. Once we know $f(x)$ we can easily compute statistically interesting quantities, like the *average* and the *variance*, which work out to be $\mu = f'(1)/f(1)$ and $\sigma^2 = f''(1)/f(1) + \mu - \mu^2$, respectively.

The general scenario is that we have an *interesting* (finite or infinite) combinatorial set, let us call it A , and a certain numerical *attribute*, $\alpha : A \rightarrow \mathbb{N}$, which assigns to each element of A a natural number. (Here we allow 0 as a natural number.) Then the *weight enumerator* of A with respect to α is defined by the formula

$$f(x) = \sum_{a \in A} x^{\alpha(a)}.$$

We shall also use the notation $|A|_x$ for $f(x)$. Obviously, this equals

$$\sum_{n=0}^{\infty} a_n x^n,$$

where a_n is the number of members of A whose α equals n . Hence if we have some kind of explicit expression for $f(x)$, we immediately have an “explicit” expression for the actual sequence a_n assuming, that is, that one considers the operations needed to calculate the n th coefficient a_n of $f(x)$ as constituting an explicit expression for a_n . Even if one does not, then it is still often possible to get a “nice” formula for a_n , or, failing this, to extract the asymptotics.

The fundamental operations for naive counting also hold for *weighted counting*: just replace $|\cdot|$ by $|\cdot|_x$. For example,

$$|A \cup B|_x = |A|_x + |B|_x$$

(if $A \cap B = \emptyset$) and

$$|A \times B|_x = |A|_x \cdot |B|_x.$$

Let us quickly see why the second of these is true. If the members of A and B are endowed with numerical attributes α and β , respectively, and one defines an attribute γ on $A \times B$ by letting $\gamma(a, b)$ equal $\alpha(a) + \beta(b)$, then

$$\begin{aligned} |A \times B|_x &= \sum_{(a,b) \in A \times B} x^{\gamma(a,b)} \\ &= \sum_{(a,b) \in A \times B} x^{\alpha(a) + \beta(b)} \\ &= \sum_{(a,b) \in A \times B} x^{\alpha(a)} \cdot x^{\beta(b)} \\ &= \sum_{a \in A} \sum_{b \in B} x^{\alpha(a)} \cdot x^{\beta(b)} \\ &= \left(\sum_{a \in A} x^{\alpha(a)} \right) \cdot \left(\sum_{b \in B} x^{\beta(b)} \right) \\ &= |A|_x \cdot |B|_x. \end{aligned}$$

Let us see how these facts can be useful. First, consider the *infinite* set A , of all (finite) sequences of 1s and 2s, and let the attribute be “sum of entries.” Then the weight of 1221 is x^6 , and, in general, the weight of a sequence $(a_1 \cdots a_r)$ is $x^{a_1 + \cdots + a_r}$. The set A can be naturally decomposed as

$$A = \{\phi\} \cup 1A \cup 2A,$$

where ϕ is the empty word, and $1A$ is short for the set of all sequences obtained by prefixing a 1 to members of A , and analogously for $2A$. Applying $|\cdot|_x$, we get

$$|A|_x = 1 + x|A|_x + x^2|A|_x,$$

which, in this simple case, can be solved *explicitly*, to yield, once again

$$|A|_x = \frac{1}{1 - x - x^2}.$$

A legal bracketing L is either empty (in which case the weight is $x^0 = 1$), or else, as we have already noted, it can be written as $L = [L_1]L_2$, where L_1 and L_2 are (shorter) legal bracketings. Conversely, whenever L_1 and L_2 are legal bracketings, so is $[L_1]L_2$. Let \mathcal{L} be the (infinite) set of *all* legal bracketings, and define the weight of a legal bracketing to be x^n , where n is the number of bracket pairs $[\]$. For example, the weight of $[\]$ is x and the weight of $[[\]][\]$ is x^5 . The set \mathcal{L} decomposes naturally as follows:

$$\mathcal{L} = \{\phi\} \cup ([\mathcal{L}] \times \mathcal{L}),$$

where ϕ denotes the empty word and $[\mathcal{L}] \times \mathcal{L}$ denotes the set of all words of the form $[L_1]L_2$ with L_1 and L_2 in \mathcal{L} . This leads to the *nonlinear* (in fact, quadratic) equation

$$|\mathcal{L}|_x = 1 + x|\mathcal{L}|_x^2,$$

which yields, thanks to the Babylonians, the explicit expression

$$|\mathcal{L}|_x = \frac{1 - \sqrt{1 - 4x}}{2x}.$$

This in turn gives us the answer to example (iii) above, via Newton's binomial theorem.

Legal bracketings are equivalent to so-called *binary trees*, that is, unlabeled ordered trees where every vertex has either no children or exactly two children. For instance, when we write the legal bracketing $[[[]][[]][[]][[]]]$ in the form $[L_1]L_2$ we can think of $[[[]][[]][[]][[]]]$ as the parent, with children $L_1 = [[]]$ and $L_2 = [[[]][[]][[]]]$. Then L_1 's children are ϕ and $[]$, while L_2 's are $[]$ and $[[[]]]$. This process continues until we have reached ϕ down every branch of the family.

If we try to count *penta-trees* instead, where each vertex may only have exactly zero or five children, then the generating function, alias weight-enumerator, satisfies the quintic equation

$$f = x + f^5,$$

which, according to ABEL [VI.33] and GALOIS [VI.41], is not *solvable by radicals* (see THE INSOLUBILITY OF THE QUINTIC [V.24]). However, solvability by radicals is not everything. More than 200 years ago, LAGRANGE [VI.22] devised a beautiful and extremely useful formula for extracting the coefficients of the generating function from the equation it satisfies, now called the *Lagrange inversion formula*. Using it one can easily show that the number of complete k -ary trees with $(k-1)m+1$ leaves is

$$\frac{(km)!}{((k-1)m+1)!m!}.$$

A multivariate generalization of the Lagrange inversion formula, discovered by the great Bayesian probabilist I. J. Good, enables one to enumerate *colored* trees and many other extensions.

3.1 Enumeration Ansatzes

If one wants to turn enumerative combinatorics into a *theory* rather than a collection of solved problems, one needs to introduce *classification*, and *enumeration paradigms* for counting sequences. But since “paradigm” is such a pretentious word, let us use the much humbler German word “ansatz,” which roughly means “form of solution.”

Let $(a_n)_{n=0}^\infty$ be a sequence, and let

$$f(x) = \sum_{n=0}^{\infty} a_n x^n$$

be its generating function. If we know the “form” of a_n , we can often deduce the form of $f(x)$ (and vice versa).

- (i) If a_n is a polynomial in n , then $f(x)$ has the form

$$f(x) = \frac{P(x)}{(1-x)^{d+1}},$$

where P is a polynomial function and d is the degree of the polynomial that describes a_n .

- (ii) If a_n is a *quasi-polynomial* in n (i.e., there exists an integer N such that for each $r = 0, \dots, N-1$, the function $m \mapsto a_{mN+r}$ is a polynomial in m), then, for some (finite) sequence of integers d_1, d_2, \dots and some polynomial function P ,

$$f(x) = \frac{P(x)}{(1-x)^{d_1}(1-x^2)^{d_2}(1-x^3)^{d_3} \dots}.$$

- (iii) If a_n is *C-recursive*, that is, if it satisfies a linear recurrence equation with constant coefficients

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + \dots + c_d a_{n-d}$$

(a good example is the Fibonacci sequence), then $f(x)$ is a *rational* function of x : that is, $f(x) = P(x)/Q(x)$, where P and Q are polynomials.

- (iv) If a_n satisfies a linear recurrence equation of the form

$$c_0(n)a_n = c_1(n)a_{n-1} + c_2(n)a_{n-2} + \dots + c_d(n)a_{n-d},$$

where the coefficients $c_i(n)$ are polynomial in n , then it is said to be *P-recursive*. (For example, $a_n = n!$ is P-recursive since we have the recurrence $a_n = na_{n-1}$.) If this is the case, then $f(x)$ is *D-finite*, which means that it satisfies a linear differential equation with polynomial coefficients (in x).

In the case of $a_n = n!$ the recurrence $a_n = na_{n-1}$ is *first order*. A natural example of a P-recursive sequence satisfying a higher-order linear recurrence with polynomial coefficients is the sequence that counts the number of involutions on $\{1, \dots, n\}$. (An involution is a permutation that equals its inverse.) Let us call this number w_n . The sequence (w_n) satisfies the recurrence relation

$$w_n = w_{n-1} + (n-1)w_{n-2}.$$

This recurrence follows from the fact that in the permutation n belongs either to a 1-cycle or to a 2-cycle. The former case accounts for w_{n-1} of the involutions, and the latter for $(n-1)w_{n-2}$ of them. (There are $n-1$ ways of choosing the cycle-mate, i , say, of n , and deleting the resulting cycle leaves an involution of the $n-2$ elements $\{1, \dots, i-1, i+1, \dots, n-1\}$.)

4 Bijective Methods

This last argument was a simple example of a *bijective proof*, in this case, of a recurrence for the number of involutions on n objects. Contrast it with the following proof.

The number of involutions of $\{1, \dots, n\}$ with exactly k 2-cycles is

$$\binom{n}{2k} \frac{(2k)!}{k!2^k},$$

because we must first choose the $2k$ elements that will participate in the k 2-cycles, and then match them up into (unordered) pairs, which can be done in

$$(2k-1)(2k-3) \cdots 1 = \frac{(2k)!}{k!2^k}$$

ways. Hence

$$w_n = \sum_k \binom{n}{2k} \frac{(2k)!}{k!2^k}.$$

Nowadays such sums can be handled completely *automatically*, and if one inputs this sum to the Maple package EKHAD (downloadable from my Web site), one would get the recurrence $w_n = w_{n-1} + (n-1)w_{n-2}$ as the output, together with a (completely rigorous!) proof. While the so-called Wilf-Zeilberger (WZ) method can handle many such problems, there are many other cases where one still needs a human proof. In either case such proofs involve (algebraic, and sometimes analytic) *manipulations*. The great combinatorialist Adriano Garsia derogatorily calls such proofs “manipulatorics,” and *real enumerators do not manipulate*, or at least try to avoid it whenever possible. The preferred method of proof is by BIJECTION [I.2 §2.2].

Suppose one has to prove that $|A_n| = |B_n|$ for every n , where A_n and B_n are combinatorial families. The “ugly way” is to get, by some means or other, algebraic or analytic expressions for $a_n = |A_n|$ and $b_n = |B_n|$. Then one *manipulates* a_n , getting another expression a'_n , which in turn leads to yet another expression a''_n , and if one is patient enough, and clever enough, and in luck, or if the problem is not too deep, one eventually arrives at b_n , and the result follows.

On the other hand, the *nice* way of proving that $|A_n| = |B_n|$ is by constructing (a preferably nice) *bijection* $T_n : A_n \rightarrow B_n$, which immediately implies, as a corollary, that $|A_n| = |B_n|$.

In addition to being more *aesthetically* pleasing, a bijective proof is also *philosophically* more satisfactory. In fact, the notion of (cardinal) *number* is a highly sophisticated *derived* notion based on the much more basic notion of *being in bijection*. Indeed, according

to FREGE [VI.56], the cardinal numbers are *equivalence classes*, where the EQUIVALENCE RELATION [I.2 §2.3] is “is in bijective correspondence with.” Saharon Shelah said that people have been exchanging objects, in a one-to-one way, since long before they started to count. Also, a bijective proof *explains why* the two sets are equinumerous, as opposed to just certifying the formal correctness of this fact.

For example, suppose that Noah had wanted to prove that there were as many male as female creatures in his Ark. One way of proving this would have been to count the males and count the females, and check that the two resulting numbers were indeed the same. But a much better, conceptual, proof would have been to note that there is an obvious one-to-one correspondence between the set M of males and the set F of females: the function $w : M \rightarrow F$ defined by $w(x) = \text{WifeOf}(x)$ is a bijection, with inverse $h : F \rightarrow M$ defined by $h(y) = \text{HusbandOf}(y)$.

A classic example of a bijective proof is Glaisher's proof of EULER's [VI.19] “odd equals distinct” partition theorem. A *partition* of an integer n is a way of writing it as a sum of positive integers, where order does not matter. For example, 6 has eleven partitions: 6, 51, 42, 411, 33, 321, 3111, 222, 2211, 21111, 111111. (Here 3111 is shorthand for the sum $3 + 1 + 1 + 1$, and so on. Since order does not matter, we count 3111 as the same partition of 6 as 1311, 1131, and 1113. It is convenient to write the partitions with their numbers in decreasing order, as we have done.)

A partition is called *odd* if all its parts are odd, and it is called *distinct* if all its parts are distinct. Let $\text{Odd}(n)$ and $\text{Dis}(n)$ be the sets of odd and distinct partitions of n , respectively. For example, $\text{Odd}(6) = \{51, 33, 3111, 111111\}$ and $\text{Dis}(6) = \{6, 51, 42, 321\}$. Euler proved that $|\text{Odd}(n)| = |\text{Dis}(n)|$ for all n . His “manipulatorics” proof goes as follows. Let $o(n)$ and $d(n)$ be the number of odd and distinct partitions of n , respectively, and let us define the *generating functions* $f(q) = \sum_{n=0}^{\infty} o(n)q^n$ and $g(q) = \sum_{n=0}^{\infty} d(n)q^n$. Using the “multiplication principle” for weighted counting, Euler showed that

$$f(q) = \prod_{i=0}^{\infty} \frac{1}{1 - q^{2i+1}} \quad \text{and} \quad g(q) = \prod_{i=0}^{\infty} (1 + q^i).$$

Using the algebraic identity $1 + y = (1 - y^2)/(1 - y)$, we have

$$\begin{aligned} \prod_{i=0}^{\infty} (1 + q^i) &= \prod_{i=0}^{\infty} \frac{1 - q^{2i}}{1 - q^i} \\ &= \frac{\prod_{i=0}^{\infty} (1 - q^{2i})}{\prod_{i=0}^{\infty} (1 - q^{2i}) \prod_{i=0}^{\infty} (1 - q^{2i+1})} \\ &= \prod_{i=0}^{\infty} \frac{1}{1 - q^{2i+1}}. \end{aligned}$$

Hence $g(q) = f(q)$, and the identity $o(n) = d(n)$ follows by extracting the coefficient of q^n .

For a very long time, these kinds of manipulation were considered to belong to the realm of *analysis*, and in order to justify the manipulations of the infinite series and products, one talked about the “region of convergence,” usually $|q| < 1$, and every step had to be justified by the appropriate analytical theorem. Only relatively recently did people come to realize that no analysis need be involved: everything makes sense in the *completely elementary* and much more rigorous (from the philosophical viewpoint) algebra of *formal power series*. One still needs to worry about convergence, so as to exclude, for example, an infinite product like $\prod_{i=0}^{\infty} (1 + x)$, but the notion of convergence in the ring of formal power series is much more user-friendly than its analytical namesake.

Even though invoking analysis was a red herring, Euler’s proof, while purely algebraic and elementary, is nevertheless still manipulatorics. It would be much nicer to find a direct bijection between the sets $\text{Dis}(n)$ and $\text{Odd}(n)$. Such a bijection was given by Glaisher. Given a distinct partition, write each of its parts as $2^r \cdot s$, where s is odd, and replace it by 2^r copies of s . (For example, $12 = 4 \cdot 3$, so we would replace 12 by $3 + 3 + 3 + 3$.) The output is obviously a partition of the same integer n , but now into odd parts. For example, the partition $(10, 5, 4)$ is transformed to the new partition $(5, 5, 5, 1, 1, 1, 1)$. To define the inverse transformation, take an odd part a and count how many times it shows up. If it shows up m times, then write m in binary notation, $m = 2^{s_1} + \dots + 2^{s_k}$, and replace the m copies of a by the k parts: $2^{s_1}a, \dots, 2^{s_k}a$. It is not hard to check that if you do the first transformation to a partition in $\text{Dis}(n)$ and then do the second transformation, you get back to the partition you started with.

When we perform algebraic (and logical, and even analytical) manipulations, we are really rearranging and combining symbols, and hence we are doing combinatorics in disguise. In fact, *everything is combinatorics*.

All we need to do is to take the combinatorics out of the closet, and make it explicit. The plus sign turns into (disjoint) union, the multiplication sign becomes Cartesian product, and induction turns into recursion. But what about the combinatorial counterpart of the minus sign? In 1982, Garsia and Steven Milne filled this gap by producing an ingenious “involution principle” that enables one to translate the implication

$$a = b \quad \text{and} \quad c = d \quad \Rightarrow \quad a - c = b - d$$

into a bijective argument, in the sense that if $C \subset A$ and $D \subset B$, and there are natural bijections $f : A \rightarrow B$ and $g : C \rightarrow D$ establishing that $|A| = |B|$, and $|C| = |D|$, then it is possible to construct an explicit bijection between $A \setminus C$ and $B \setminus D$. Let us define it in terms of people. Suppose that in a certain village all the adults are married, with the result that there is a natural bijection from the set of married men to the set of married women, $m \mapsto \text{WifeOf}(m)$, with its inverse $w \mapsto \text{HusbandOf}(w)$. In addition, some of the people have extramarital affairs, but only one per person, and all within the village. There is a natural bijection from the set of cheating men to the set of cheating women, called $m \mapsto \text{MistressOf}(m)$, with its inverse $w \mapsto \text{LoverOf}(w)$. It follows that there are as many faithful men as there are faithful women. But how do we match them up? (One might imagine, for example, that each faithful man wants a faithful woman to go to church with him.)

Here is how it is done. A faithful man first asks his wife to come with him. If she is faithful, she agrees. If she is not, she has a lover, and that lover has a wife. So she tells her husband: “Sorry, hubby, I am going to the pub with my lover, but my lover’s wife may be free.” If this happens, then the man asks the wife of the lover of his wife to go with him, and if she is faithful, she agrees. If she is not he keeps asking the wife of the lover of the woman who has just rejected his proposal. Since the village is finite, he will eventually get to a faithful woman.

The reaction of the combinatorial enumeration community to the involution principle was mixed. On the one hand it had the universal appeal of a general principle, one that should be useful in many attempts to find bijective proofs of combinatorial identities. On the other hand, its universality is also a major drawback, since involution-principle proofs usually do not give any insight into the *specific* structures involved, and one feels a bit cheated. Such a proof answers the *letter* of the question, but it misses its *spirit*. Given a proof of this kind, one still hopes for a *really* natural,

“involution-principle-free proof.” This is the case, for instance, with the celebrated Rogers-Ramanujan identity, which states that the number of partitions of an integer into parts that leave remainder 1 or 4 when divided by 5 equals the number of partitions of that integer with the property that the difference between any two parts is at least 2. For example, if $n = 7$ the cardinalities of $\{61, 4111, 1111111\}$ and $\{7, 61, 52\}$ are the same. Garsia and Milne invented their notorious principle in order to give a Rogers-Ramanujan bijection, thereby winning a \$50 prize from George Andrews. However, finding a *really nice* bijective proof is still an open problem.

A quintessential example of a bijective proof is Prüfer’s proof of CAYLEY’S [VI.46] celebrated result that there are n^{n-2} labeled trees on n vertices (example (v) earlier). Recall that a labeled tree is a labeled connected simple graph without cycles. Every tree has at least two vertices with only one neighbor (these are called *leaves*). A certain mapping called the *Prüfer bijection* associates with every labeled tree T a vector of integers (a_1, \dots, a_{n-2}) , with $1 \leq a_i \leq n$ for each i . This vector is called its *Prüfer code*. Since there are n^{n-2} such vectors, Cayley’s formula follows once we have defined the mapping $f : \text{Trees} \rightarrow \text{Codes}$ and proved that it is indeed a bijection. This really needs four steps: defining f , defining its alleged inverse map g , and proving that $g \circ f$ and $f \circ g$ are the identity maps on their respective domains.

The mapping f is defined recursively as follows. If the tree has 2 vertices, then its code is the empty sequence. Otherwise, let a_1 be the (sole) neighbor of the smallest leaf and let (a_2, \dots, a_{n-2}) be the code of the smaller tree obtained by deleting that leaf.

5 Exponential Generating Functions

So far, when we have discussed generating functions, we have been talking about *ordinary generating functions* (or OGFs). These are ideally suited for counting ordered structures like integer partitions, ordered trees, and words. But many combinatorial families are really *sets*, where the order is immaterial. For these the natural concept is that of an *exponential generating function* (or EGF).

The EGF of a sequence $\{a(n)\}_{n=0}^{\infty}$ is defined to be

$$\sum_{n=0}^{\infty} \frac{a(n)}{n!} x^n.$$

Labeled objects can be often viewed as sets of smaller *irreducible* objects. For example, a permutation is the

disjoint union of *cycles*, a set partition is the disjoint union of *nonempty sets*, a (labeled) forest is the disjoint union of *labeled trees*, and so on.

Suppose that we have two combinatorial families A and B , and suppose that there are $a(n)$ labeled objects of size n in the A family, and $b(n)$ in the B family. We can construct a new set of labeled objects $C = A \times B$, where the labels are disjoint and distinct, and define the size of a pair to be the sum of the sizes of the components. We have

$$c(n) = \sum_{k=0}^n \binom{n}{k} a(k) b(n-k),$$

since we must

- (i) decide the size of the first component, k (an integer between 0 and n), which forces the size of the second component to be $n - k$,
- (ii) decide which of the n labels go to the first component ($\binom{n}{k}$ ways), and
- (iii) pick the objects for each component from the A and B families, respectively, using the available labels ($a(k)b(n-k)$ ways).

Multiplying both sides by $x^n/n!$ and summing from $n = 0$ to $n = \infty$ yields

$$\begin{aligned} \sum_{n=0}^{\infty} \frac{c(n)}{n!} x^n &= \sum_{n=0}^{\infty} \sum_{k=0}^n \frac{a(k)}{k!} x^k \frac{b(n-k)}{(n-k)!} x^{n-k} \\ &= \left(\sum_{k=0}^{\infty} \frac{a(k)}{k!} x^k \right) \left(\sum_{n-k=0}^{\infty} \frac{b(n-k)}{(n-k)!} x^{n-k} \right). \end{aligned}$$

Hence $\text{EGF}(C) = \text{EGF}(A) \text{EGF}(B)$. Iterating, we get

$$\text{EGF}(A_1 \times A_2 \times \cdots \times A_k) = \text{EGF}(A_1) \cdots \text{EGF}(A_k).$$

In particular, if all the A_i are the same, we have that the EGF of ordered k -tuples, A^k , equals $[\text{EGF}(A)]^k$. But if “order does not matter,” then the EGF of k -sets of A -objects is $[\text{EGF}(A)]^k/k!$, since there are exactly $k!$ ways of arranging a k -set into an ordered array (since all labels are distinct, all these objects are different). Summing from $k = 0$ to $k = \infty$ we get the “fundamental theorem of exponential generating functions.”

If B is a labeled combinatorial family that can be viewed as sets of “connected components” that belong to a combinatorial family A , then

$$\text{EGF}(B) = \exp[\text{EGF}(A)].$$

This useful theorem was part of the physics folklore for many years, and was also implicit in many older combinatorial proofs. However, it was explicated only

in the early 1970s. It was fully “categorized” by means of Joyal’s theory of species, which grew to be a beautiful theory of enumeration in the hands of the *école Québécoise* (the Labelle and Bergeron frères, Leroux, and others).

Here are some venerable examples. Let us try to find the EGF of set partitions. That is, let us try to figure out an expression for

$$\sum_{n=0}^{\infty} \frac{b(n)}{n!} x^n,$$

where $b(n)$ (so-called Bell numbers) denotes the number of set partitions of an n -element set.

Recall that a *set partition* of a set A is a set of pairwise-disjoint *nonempty* subsets of A , $\{A_1, \dots, A_r\}$, such that the union of all the A_i equals A . For example, the set partitions of the 2-element set $\{1, 2\}$ are $\{\{1\}, \{2\}\}$ and $\{\{1, 2\}\}$.

The atomic objects in this example are *nonempty sets*. (We think of a set A as being the “trivial” partition of itself into just one set.) Let $a(n)$ be the number of ways of partitioning a set of size n into one nonempty set. Clearly, when $n = 0$ this cannot be done, so $a(0) = 0$. When $n = 1$ there is exactly one way of doing it, so the EGF of the sequence $a(n)$ is

$$A(x) = 0 + \sum_{n=1}^{\infty} \frac{1}{n!} x^n = e^x - 1.$$

It follows immediately from the fundamental theorem that

$$\sum_{n=0}^{\infty} \frac{b(n)}{n!} x^n = e^{e^x - 1}, \quad (1)$$

an identity of Bell. Nowadays, with computer algebra systems, this can be used immediately to crank out the first 100 terms of the sequence $b(n)$. For example, in Maple one simply types

```
taylor(exp(exp(x))-1, x=0, 101);
```

so this is definitely an answer in the Wilfian sense. We can also easily derive *recurrences* (albeit ones that need at least $O(n)$ memory), by differentiating both sides of (1) and comparing coefficients.

That was really easy, so let us go on and prove something much deeper. How about an EGF-style proof of Levi Ben Gerson’s celebrated formula for the number of permutations on n objects, $n!$ (example (ii) earlier)? Every permutation can be decomposed into a disjoint union of cycles, so the atomic objects are now *cycles*. How many n -cycles are there? The answer is of course $(n-1)!$, since (a_1, a_2, \dots, a_n) is the

same as $(a_2, a_3, \dots, a_n, a_1)$, which is the same as $(a_3, \dots, a_n, a_1, a_2)$, etc., which means that we can pick the first entry arbitrarily, after which we have $(n-1)!$ choices for placing the remaining entries. The EGF for cycles is therefore

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{(n-1)!}{n!} x^n &= \sum_{n=1}^{\infty} \frac{1}{n} x^n \\ &= -\log(1-x) = \log(1-x)^{-1}. \end{aligned}$$

Using the fundamental theorem, we get that the EGF of permutations is

$$\exp(\log(1-x)^{-1}) = (1-x)^{-1} = \sum_{n=0}^{\infty} x^n = \sum_{n=0}^{\infty} \frac{n!}{n!} x^n,$$

and voilà we have a beautiful new proof that the number of permutations on n objects is $n!$.

This argument may not look very impressive. But a slight modification leads immediately to the (ordinary) generating function for the number of permutations on $\{1, \dots, n\}$ with exactly k cycles, which we shall denote by $c(n, k)$. Here we are fixing n and letting k vary, so the generating function is $C_n(\alpha) = \sum_{k=0}^n c(n, k) \alpha^k$. All we have to do to calculate this is go from *naïve* counting to *weighted* counting, and assign to each permutation the weight $\alpha^{\# \text{cycles}}$. The fundamental theorem of exponential generating functions carries over word-for-word to weighted counting. The weighted EGF for cycles is $\alpha \log(1-x)^{-1}$, so the weighted EGF for permutations is

$$\exp(\alpha \cdot \log(1-x)^{-1}) = (1-x)^{-\alpha} = \sum_{n=0}^{\infty} \frac{(\alpha)_n}{n!} x^n,$$

where

$$(\alpha)_n = \alpha(\alpha+1) \cdots (\alpha+n-1)$$

is the so-called *rising factorial*. We have therefore derived the far less trivial result that the number of permutations of $\{1, \dots, n\}$ with exactly k cycles equals the coefficient of α^k in $(\alpha)_n$.

About ten years ago (Ehrenpreis and Zeilberger 1994) I used this technique to give a combinatorial proof of the Pythagorean theorem in the form

$$\sin^2 z + \cos^2 z = 1.$$

The functions $\sin z$ and $\cos z$ are the weighted EGFs for *increasing sequences* of odd and even lengths, respectively, with weight $(-1)^{\lfloor \text{length}/2 \rfloor}$. Hence the left-hand side is the weighted EGF for ordered pairs of increasing sequences

$$a_1 < \cdots < a_k, \quad b_1 < \cdots < b_r,$$

T&T: check double space not deleted here before CRC.

such that k and r have the same parity, the sets $\{a_1, \dots, a_k\}$ and $\{b_1, \dots, b_r\}$ are disjoint, and the union of the two sets is $\{1, 2, \dots, k + r\}$. There is a killer involution on these sets of pairs defined as follows.

If $a_k < b_r$ then map the pair to

$$a_1 < \dots < a_k < b_r, \quad b_1 < \dots < b_{r-1}.$$

and otherwise map it to

$$a_1 < \dots < a_{k-1}, \quad b_1 < \dots < b_r < a_k.$$

For example, the pair

$$1, 3, 5, 6 \quad 2, 4, 7, 8, 9, 10, 11, 12,$$

whose sign is $(-1)^2 \cdot (-1)^4 = 1$, goes to the pair

$$1, 3, 5, 6, 12 \quad 2, 4, 7, 8, 9, 10, 11,$$

whose sign is $(-1)^2 \cdot (-1)^3 = -1$ (and vice versa).

Since this mapping changes the sign, and is an involution, all such pairs can be paired up into mutually canceling pairs. But this mapping is undefined for one special pair, namely the pair (empty, empty), whose weight is 1. Therefore, the EGF for the sum of the weights of all pairs is 1, which explains the right-hand side.

Yet another application of this method is a proof of André's generating function for the number of *up-down* permutations. A permutation of $a_1 \dots a_n$ is called up-down (or sometimes *zigzag*) if $a_1 < a_2 > a_3 < a_4 > a_5 < \dots$. Let a_n be the number of up-down permutations. Then

$$\sum_{n=0}^{\infty} \frac{a(n)}{n!} x^n = \sec x + \tan x.$$

This is equivalent to saying that

$$\cos x \cdot \left(\sum_{n=0}^{\infty} \frac{a(n)}{n!} x^n \right) = 1 + \sin x.$$

Can you find the appropriate set and the killer involution?

6 Pólya-Redfield Enumeration

Often in enumeration it is easy enough to count *labeled* objects, but what about unlabeled ones? For example, the number of labeled (simple) graphs on n vertices (example (vi)) is trivially $2^{n(n-1)/2}$, but how many unlabeled graphs are there on n vertices? This is much harder, and in general there are no "nice" answers, but the best known way is via a powerful technique initiated by Pólya, which was largely anticipated by Redfield. Pólya enumeration lends itself very efficiently to

counting chemical isomers, since, for example, all the carbon atoms "look the same." Indeed, counting isomers was Pólya's initial motivation (see MATHEMATICS AND CHEMISTRY [VII.1 §2.3]).

The main idea is to view *unlabeled* objects as equivalence classes of easy-to-count *labeled* objects, and to count these equivalence classes. But what is the equivalence? The answer is that there is always a SYMMETRY GROUP [L3 §2.1] involved, and it leads to a natural equivalence relation. Let the symmetry group be G , and let the set of labeled objects be A . Then two objects a and b of A are regarded as *equivalent* if $b = g(a)$ for some member g of the group G . This means that there is some symmetry g in the group G that transforms a to b . This is easily seen to be an equivalence relation and the equivalence classes are the sets

$$\text{Orbit}(a) = \{g(a) \mid g \in G\}, \quad a \in A,$$

which are known as *orbits*. Calling each orbit a "family," we have the task of counting the number of families. Note that G is a subgroup of the group of permutations of the finite set A .

Suppose that there is a picnic consisting of many families and we want to count the number of families. One way would be to define some "canonical head" of each family, say "mother," and count the number of mothers. But some daughters look like mothers, so this is not so easy. On the other hand, you cannot just count everybody, since then you would count each family several times. The problem is that "naive" counting of people (or objects) is giving a credit of 1 to each person, and this is inappropriate if we are trying to count families. If instead we were to ask each person "How big is your family?" and add to our count the reciprocal of that number, then the calculation would come out just right, since a family of size k would get a credit of $1/k$ for each of its members, and would therefore have been counted exactly once by the end. Going back to counting orbits, we see by the same reasoning that their number is

$$\sum_{a \in A} \frac{1}{|\text{Orbit}(a)|}.$$

The conceptual opposite of "orbit of a " is the subgroup of members of G that fix a :

$$\text{Fix}(a) = \{g \in G \mid g(a) = a\}.$$

(This is sometimes known as the *stabilizer* of a .) To each element $b = ga$ in the orbit of a , we can associate the left coset $g\text{Fix}(a)$ of $\text{Fix}(a)$. This association

turns out to be a well-defined one-to-one correspondence between the orbit of a and the cosets of $\text{Fix}(a)$ in G , from which it follows that the size of $\text{Orbit}(a)$ is $|G/\text{Fix}(a)|$. We can therefore substitute $|\text{Fix}(a)|/|G|$ for $1/|\text{Orbit}(a)|$ in the previous formula, which implies that the number of orbits is

$$\frac{1}{|G|} \sum_{a \in A} |\text{Fix}(a)|.$$

Let us use the notation $\chi(\text{statement})$ to stand for 1 if the statement is true and 0 if it is false. Then

$$\begin{aligned} \frac{1}{|G|} \sum_{a \in A} |\text{Fix}(a)| &= \frac{1}{|G|} \sum_{a \in A} \sum_{g \in G} \chi(g(a) = a) \\ &= \frac{1}{|G|} \sum_{g \in G} \sum_{a \in A} \chi(g(a) = a) \\ &= \frac{1}{|G|} \sum_{g \in G} \text{fix}(g), \end{aligned}$$

where $\text{fix}(g)$ is the number of fixed points of g (when g is viewed as a permutation of A). We have just proved what used to be called *Burnside's lemma*, but it goes back to CAUCHY [VI.29] and FROBENIUS [VI.58]. It states that the total number of orbits equals the average number of fixed points of g , over all transformations g in G . If the group G is the full symmetric group of all the permutations of A , then the average number of fixed points equals 1 (since in this trivial case there is only one orbit!).

Enter Pólya. The objects that he was interested in counting (e.g., chemical isomers, or colorings of the faces of the cube) were all naturally *functions* from an *underlying set* to a set of *colors* (or atoms). Let us call the underlying set U and the set of colors C . A symmetry of U gives rise in a natural way to a transformation of the set of functions $f : U \rightarrow C$. Given a function f one defines a new function gf by $g(f)(u) = f(g(u))$. (If we think of f as a coloring, then gf is the new coloring that assigns to u the color that f assigned to $g(u)$.) Now let us think about the number of fixed points of g in the set of C -colorings of U . Such a fixed point is a coloring f that equals gf : that is, $f(u) = f(gu)$ for every u . But then $f(u) = f(gu) = f(g^2u) = \dots$, which means that, given any cycle of g , f must assign the same color to all members of that cycle. It follows that the number of fixed colorings of g is $c^{\#\text{cycles}(g)}$, where $c = |C|$ is the number of colors.

Applying Burnside's lemma, we may deduce that the number of different colorings of U (up to G -equivalence) is

$$\frac{1}{|G|} \sum_{g \in G} c^{\#\text{cycles}(g)},$$

since an equivalence class of colorings is simply an orbit of one of the colorings in that class.

Here is a simple application. How many necklaces (without a clasp) are there that consist of p beads (where p is a prime) and that use a different colors? The underlying set is $\{0, \dots, p-1\}$, and the symmetry group is \mathbb{Z}_p , the cyclic group of order p . As usual, regard the elements of the symmetry group as permutations of the set of beads. Since p is a prime, there are $p-1$ elements of \mathbb{Z}_p with one cycle (of length p), and one element (the identity permutation) with p cycles (all of length 1). It follows that the number of necklaces is

$$\frac{1}{p} ((p-1) \cdot a + 1 \cdot a^p) = a + \frac{a^p - a}{p}.$$

In particular, since this number is necessarily an integer, we get as a bonus a combinatorial proof of FERMAT'S LITTLE THEOREM [III.60]: that $a^p - a$ is always a multiple of p . Perhaps one day there will be an equally nice combinatorial proof of Fermat's *last* theorem. All one has to do is to prove that there is no bijection from the union of the set of straight necklaces of size n using x colors, and the set of such necklaces using y colors, to the set of necklaces using z colors (with $n > 2$, of course).

If one wants to keep track of how many beads there are of each color, one simply replaces straight counting by weighted counting, and $c^{\#\text{cycles}(g)}$ is replaced by

$$(x_1 + \dots + x_c)^{\alpha_1} \cdot (x_1^2 + \dots + x_c^2)^{\alpha_2} \dots$$

(assuming that g has α_1 1-cycles, α_2 2-cycles, etc.). The resulting expression is the celebrated *cycle-index polynomial*.

6.1 The Principle of Inclusion-Exclusion and Möbius Inversion

Another pillar of enumeration is the principle of inclusion-exclusion (nicknamed PIE). Suppose that there are n sins, s_1, \dots, s_n , that a person may succumb to, and suppose that for each set of sins S , A_S is the set of people who have all the sins in S (and possibly others). Then the number of good people (without sins) is

$$\sum_S (-1)^{|S|} |A_S|.$$

For example, if the set A is the set of all permutations π of $\{1, \dots, n\}$ and the i th sin is having $\pi[i] = i$, then $|A_S| = (n - |S|)!$, and we get that the number of *derangements* (permutations without fixed points) is

$$\sum_{k=0}^n (-1)^k \binom{n}{k} (n-k)! = n! \sum_{k=0}^n (-1)^k \frac{1}{k!},$$

which yields the *answer*: “closest integer to $n!/e$.” This is sometimes called the “umbrella problem”: if on a rainy day n absent-minded people go to a party and leave an umbrella by the door, and if on their departure they each take a random umbrella, then the probability that nobody ends up with the right umbrella is about $1/e$.

The PIE is a special case of *Möbius inversion* on general partially ordered sets (posets) where the poset happens to be the Boolean lattice. This realization was published in a seminal paper by Rota (1964) and reprinted in his collected works. It is considered by many to be the big bang that started modern algebraic combinatorics. Möbius’s original inversion formula is recovered when the partially ordered set is \mathbb{N} and the partial order is divisibility.

A contemporary account of enumeration from the “algebraic” point of view can be found in a marvelous two-volume set by Stanley (2000), which I strongly recommend.

7 Algebraic Combinatorics

So far I have described one of the routes to algebraic combinatorics: abstraction and conceptualization of classical enumeration. The other route, “concretization of the abstract,” is almost everywhere dense in mathematics, and cannot be described in a few pages. Let me quote from the preface of the excellent *New Perspectives in Algebraic Combinatorics* by Billera et al. (1999).

Algebraic combinatorics involves the use of techniques from algebra, topology, and geometry in the solution of combinatorial problems, or the use of combinatorial methods to attack problems in these areas. Problems amenable to the methods of algebraic combinatorics arise in these or other areas of mathematics or from diverse parts of applied mathematics. Because of this interplay with many fields of mathematics, algebraic combinatorics is an area in which a wide variety of ideas and methods come together.

7.1 Tableaux

An interesting class of objects that initially came up in group representation theory, but that turned out to be useful in many other areas—such as, for example, the theory of algorithms—are *Young tableaux*. They were first used by Reverend Alfred Young to construct *explicit* bases for the IRREDUCIBLE REPRESENTATIONS [IV.9 §2] of the SYMMETRIC GROUP [III.70]. For any partition $\lambda = \lambda_1 \cdots \lambda_k$ of n , a Young tableau of shape λ is

an array of k left-justified rows with λ_1 entries in the first row, λ_2 entries in the second row, and so on, such that every row and every column is increasing, and the set of entries is $\{1, 2, \dots, n\}$. For example, there are two standard Young tableaux whose shape is 22,

$$\begin{array}{cc} 1 & 2 \\ 3 & 4 \end{array} \quad \begin{array}{cc} 1 & 3 \\ 2 & 4 \end{array},$$

and three of shape 31,

$$\begin{array}{ccc} 1 & 2 & 3 \\ 4 & & \end{array} \quad \begin{array}{ccc} 1 & 2 & 4 \\ 3 & & \end{array} \quad \begin{array}{ccc} 1 & 3 & 4 \\ 2 & & \end{array}.$$

Let f_λ be the number of standard Young tableaux of shape λ . For example, for $n = 4$: $f_4 = 1$, $f_{31} = 3$, $f_{22} = 2$, $f_{211} = 3$, and $f_{1111} = 1$. The sum of the squares of these numbers is $1^2 + 3^2 + 2^2 + 3^2 + 1^2 = 24 = 4!$.

The number f_λ is the dimension of the irreducible representation parametrized by λ . It follows by a result in REPRESENTATION THEORY [IV.9] known as *Frobenius reciprocity* that the same is true for all n . In other words,

$$\sum_{\lambda \vdash n} f_\lambda^2 = n!,$$

a result known as the *Young–Frobenius identity*. A gorgeous *bijective* proof of this identity, which has many beautiful properties, was given by Gilbert Robinson and Craige Schensted and later extended by Donald Knuth, and is now known as the Robinson–Schensted–Knuth correspondence. It inputs a permutation $\pi = \pi_1 \pi_2 \cdots \pi_n$, and outputs a pair of Young tableaux of the same shape, thereby proving the identity.

Algebraic combinatorics is currently a very active field, and as mathematics is becoming more and more concrete, constructive, and algorithmic, there are going to be many more combinatorial structures discovered in all areas of mathematics (and science!) and this will guarantee that algebraic combinatorialists will stay very busy for a long time to come.

Further Reading

- Billera, L. J., A. Björner, C. Greene, R. E. Simion, and R. P. Stanley, eds. 1999. *New Perspectives in Algebraic Combinatorics*. Cambridge: Cambridge University Press.
- Ehrenpreis, L., and D. Zeilberger. 1994. Two EZ proofs of $\sin^2 z + \cos^2 z = 1$. *American Mathematical Monthly* 101: 691.
- Rota, G.-C. 1964. On the foundations of combinatorial theory. I. Theory of Möbius functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 2:340–68.
- Stanley, R. P. 2000. *Enumerative Combinatorics*, volumes 1 and 2. Cambridge: Cambridge University Press.

IV.19 Extremal and Probabilistic Combinatorics

Noga Alon and Michael Krivelevich

1 Combinatorics: An Introduction

1.1 Examples

It is hard to give a rigorous definition of combinatorics. Instead, let us start with a few examples to illustrate what the area is about.

(i) While examining friendship between children some fifty years ago, the Hungarian sociologist Sandor Szalai observed that among any group of about twenty children he checked he could always find four children any two of whom were friends, or else four children no two of whom were friends. Despite the temptation to try to draw sociological conclusions, Szalai realized that this might well be a mathematical phenomenon rather than a sociological one. Indeed, a brief discussion with the mathematicians Erdős, Turán, and Sós convinced him this was the case. If X is any set of size 18 or more, and R is some symmetric RELATION [I.2 §2.3] on X , then there is always a subset S of X of size 4 with the following property: either xRy for any two distinct elements x, y of S , or xRy for no two distinct elements x, y of S . In this case, X is a set of children and R is the relation “is friends with.” This mathematical fact is a special case of *Ramsey’s theorem*, which was proved by the economist and mathematician Frank Plumpton Ramsey in 1930. Ramsey’s theorem led to the development of Ramsey theory, a branch of *extremal combinatorics*, which will be discussed in the next section.

(ii) In 1916, Schur was studying FERMAT’S LAST THEOREM [V.12]. It is sometimes possible to prove that a Diophantine equation has no solutions by showing that it has no solutions mod p for some prime p (see MODULAR ARITHMETIC [III.60]). However, Schur proved that for every integer k and every sufficiently large prime p , there are three integers a, b , and c , none of them congruent to 0 mod p , such that $a^k + b^k$ is congruent to c^k . Although this is a result in number theory, it has a relatively simple and purely combinatorial proof, which is another example of the many applications of Ramsey theory.

(iii) When studying the number of real zeros of random polynomials, LITTLEWOOD [VI.79] and Offord investigated in 1943 the following problem. Let $z_1, z_2,$

\dots, z_n be n not-necessarily-distinct complex numbers, each of modulus at least 1. One can form 2^n sums by taking some subset of these numbers and adding them together (with the convention that if one takes the empty set, then the sum is 0). Littlewood and Offord wanted to know how many of these sums there could conceivably be such that the difference between any two of them had modulus less than 1. When $n = 2$ the answer is easily seen to be at most 2. There are four sums: 0, z_1 , z_2 , and $z_1 + z_2$. You cannot choose both of the first two or both of the last two or you will have a difference of z_1 , which has modulus at least 1. Kleitman and Katona proved that in general the maximum is $\binom{n}{\lfloor n/2 \rfloor}$. Notice that a simple construction proves that this maximum can be achieved. Indeed, let $z_1 = z_2 = \dots = z_n$ and choose all sums of precisely $\lfloor n/2 \rfloor$ of them. There are $\binom{n}{\lfloor n/2 \rfloor}$ such sums and they are all equal. The proof that one cannot do better than this uses tools from another area of extremal combinatorics, where the basic objects studied are systems of finite sets.

(iv) Consider a school in which there are m teachers T_1, T_2, \dots, T_m and n classes C_1, C_2, \dots, C_n . The teacher T_i has to teach the class C_j for a specified number p_{ij} of lessons. What is the minimum possible number of periods in a complete timetable? Let d_i denote the total number of lessons the teacher T_i has to teach, and let c_j denote the total number of lessons the class C_j has to be taught. Clearly, the number of periods required for a complete schedule is at least as big as any d_i or c_j , and thus at least as big as the maximum of all these numbers, which we denote by d . It turns out that this obvious lower bound of d is also an upper bound: it is always possible to fit all the lessons that need to be taught into d periods. This is a consequence of *König’s theorem*, which is a basic result in graph theory. Suppose now that the situation is not so simple: for every teacher T_i and every class C_j there is some specified set of d periods in which the teaching has to take place. Can we always find a feasible timetable with these more complicated constraints? Recent breakthroughs from a subject known as *list coloring* of graphs imply that it is always possible.

(v) Given a map with several countries represented, how many colors do you need if you want to color the countries without giving any two adjacent countries the same color? Here we assume that each country forms a connected region in the plane. Of course, at least four colors may be necessary: think of Belgium, France, Germany, and Luxembourg, out of which any

T&T note: check that suboptimal linebreak does not fall at a pagebreak here before CRC.

two have a common border. The FOUR-COLOR THEOREM [V.14], proved by Appel and Haken in 1976, asserts that you never need *more* than four colors. The study of this problem led to numerous interesting questions and results about graph coloring.

(vi) Let S be an arbitrary subset of the two-dimensional lattice \mathbb{Z}^2 . For any two finite subsets $A, B \subset \mathbb{Z}$ we can think of the Cartesian product $A \times B$ as a sort of “combinatorial rectangle.” This set has size $|A| |B|$ (where $|X|$ denotes the size of a set X), and we can define an obvious notion of the *density* $d_S(A, B)$ of S in $A \times B$ by the formula $d_S(A, B) = |S \cap (A \times B)| / |A| |B|$, which measures what proportion of the elements of $A \times B$ belong to S . For each k , let $d(S, k)$ be the largest possible value of $d_S(A, B)$ if $|A| = |B| = k$. What can we say about $d(S, k)$ as k tends to infinity? One might guess that almost any behavior is possible, but, remarkably, basic results in extremal graph theory (about the so-called Turán numbers of complete bipartite graphs) imply that $d(S, k)$ must always tend to 0 or 1.

(vii) Suppose that n basketball teams compete in a tournament and any two teams play each other exactly once. The organizers wish to award k prizes at the end of the tournament. It would be embarrassing if there ended up being a team that had not won a prize despite beating all the teams that had won a prize. However, unlikely though it might sound, it is quite possible that this will be the case *whatever* k teams they choose, at least if n is large enough. To demonstrate this is easy if one uses the *probabilistic method*, which is one of the most powerful techniques in combinatorics. For any fixed k , and all sufficiently large n , if the results of all the games are chosen randomly (and uniformly and independently), then there is a very high probability that for any k teams there is another team that beats all of them. Probabilistic combinatorics, which is one of the most active areas in modern combinatorics, started with the realization that probabilistic reasoning often provides simple solutions to problems of this type, problems that are often very hard to solve in any other way.

(viii) If G is a finite group of n elements, and H is a subgroup of size k in G , then there are n/k left cosets and n/k right cosets of H . Is there always a set of n/k elements of G that contains a single representative of each right coset and a single representative of each left coset? Hall’s theorem, a basic result in graph theory, implies that there is. In fact, if H' is another subgroup of size k in G , then there is always a set of n/k elements of G that contains a single representative of each right

coset of H and a single representative of each left coset of H' . This may sound like a result in group theory, but it is really a (simple) result in combinatorics.

1.2 Topics

The examples described above illustrate some of the main themes of combinatorics. The subject, sometimes also called *discrete mathematics*, is a branch of mathematics that focuses on the study of discrete objects (as opposed to continuous ones) and their properties. Although combinatorics is probably as old as the human ability to count, the field has experienced tremendous growth during the last fifty years and has matured into a thriving area with its own set of problems, approaches, and methodology.

The examples above suggest that combinatorics is a basic mathematical discipline that plays a crucial role in the development of many other mathematical areas. In this essay we discuss some of the main aspects of this modern field, focusing on extremal and probabilistic combinatorics. (An account of combinatorial problems with a rather different flavor can be found in ALGEBRAIC AND ENUMERATIVE COMBINATORICS [IV.18].) It is, of course, impossible to cover the area fully in such a short article. A detailed account of the subject can be found in Graham, Grötschel, and Lovász (1995). Our main intention is to give a glimpse of the topics, methods, and applications illustrated by representative examples. The topics we discuss include extremal graph theory, Ramsey theory, the extremal theory of set systems, combinatorial number theory, combinatorial geometry, random graphs, and probabilistic combinatorics. The methods applied in the area include combinatorial techniques, probabilistic methods, tools from linear algebra, spectral techniques, and topological methods. We also discuss the algorithmic aspects and some of the many fascinating open problems in the area.

2 Extremal Combinatorics

Extremal combinatorics deals with the problem of determining or estimating the maximum or minimum possible size of a collection of finite objects that satisfies certain requirements. Such problems are often related to other areas, including computer science, information theory, number theory, and geometry. This branch of combinatorics has developed spectacularly over the last few decades (see, for example, Bollobás (1978), Jukna (2001), and their many references).

2.1 Extremal Graph Theory

A GRAPH [III.34] is one of the very basic combinatorial structures. It consists of a set of points, called *vertices*, some of which are linked by *edges*. One can represent a graph visually by drawing the vertices as points in the plane and the edges as lines (or curves). However, formally a graph is more abstract: it is just a set together with a collection of pairs taken from the set. More precisely, it consists of a set V , called the *vertex set*, and a set E , called the *edge set*; the elements of E (the edges) are sets of the form $\{u, v\}$, where u and v are distinct elements of V . If $\{u, v\}$ is an edge, we say that u and v are *adjacent*. The *degree* $d(v)$ of a vertex v is the number of vertices adjacent to it.

Here are a number of simple definitions associated with graphs that have emerged as important. A *path* of length k from u to v in G is a sequence of distinct vertices $u = v_0, v_1, \dots, v_k = v$, where v_i and v_{i+1} are adjacent for all $i < k$. If $v_0 = v_k$ (but all vertices v_i for $i < k$ are distinct), this is called a *cycle* of length k , and is usually denoted by C_k . A graph G is *connected* if for any two vertices u, v of G there is a path from u to v . A *complete graph* K_r is a graph with r vertices such that any two of them are adjacent. A *subgraph* of a graph G is a graph that contains some of the vertices of G and some of its edges. A *clique* in G is a set of vertices in G such that any two of them are adjacent. The maximum size of a clique in G is called the *clique number* of G . Similarly, an *independent set* in G is a set of vertices in G with *no* two of them adjacent, and the *independence number* of G is the maximum size of an independent set in it.

Extremal graph theory deals with quantitative connections between various parameters of a graph, such as its numbers of vertices and edges, its clique number, or its independence number. In many cases a certain optimization problem involving these parameters has to be solved (for example, determining how big one parameter can be if another one is at most some given size), and its optimal solutions are the *extremal graphs* for this problem. Many important optimization problems that do not explicitly mention graphs can be reformulated, using the definitions above, as problems about extremal graphs.

2.1.1 Graph Coloring

Let us return to the map-coloring example discussed in the introduction. To translate the problem into mathematics, we can describe the map-coloring problem in

terms of a graph G , as follows. The vertices of G correspond to the countries on the map, and two vertices are connected by an edge in G if and only if the corresponding countries share a common border. It is not hard to show that one can draw such a graph in such a way that no two edges cross each other: such graphs are called *planar*. Conversely, any planar graph arises in this way. Therefore, our problem is equivalent to the following: if you want to color the vertices of a planar graph so that no two adjacent vertices receive the same color, then how many colors do you need? (One can make the problem yet more mathematical by removing the nonmathematical notion of color. For example, one can assign to each vertex a positive integer instead.) Such a coloring is called *proper*. In this language, the four-color theorem states that every planar graph can be properly colored with four colors.

Here is another example of a graph-coloring problem. Suppose we must schedule meetings of several parliament committees. We do not wish to have two committees meeting at the same time if some parliament member belongs to both, so how many sessions do we need?

Again we can model this situation by using a graph G . The vertices of G represent the committees, with two vertices adjacent if and only if the corresponding committees share a member. A *schedule* is a function f that assigns to each committee one of k time slots. More mathematically, we can think of it as just a function from V to the set $\{1, 2, \dots, k\}$. Let us call a schedule *valid* if no two adjacent vertices are assigned the same number. This corresponds to no two committees being assigned the same time slot if they share a member. The question then becomes, "What is the minimal value of k for which a valid schedule exists?"

The answer is called the *chromatic number* of the graph G , denoted $\chi(G)$: it is the smallest number of colors in any proper coloring of G . Notice that a coloring of a graph G is proper if and only if for each color the set of vertices of that color is independent. Therefore, $\chi(G)$ can also be defined as the smallest number of independent sets into which it is possible to partition the vertices of G . A graph is called *k-colorable* if it admits a k -coloring, or, equivalently, if it can be partitioned into k independent sets. Thus, $\chi(G)$ is the minimum k for which G is k -colorable.

Two simple examples are in order. If G is a complete graph K_n on n vertices, then obviously in any coloring of G all vertices get distinct colors, and thus n colors are necessary. Of course, n colors are also suffi-

cient, so $\chi(K_n) = n$. If G is a cycle C_{2n+1} on $2n + 1$ vertices, then easy parity arguments show that at least three colors are needed, and three colors are enough: color the vertices along the cycle alternately by colors 1 and 2, and then color the last vertex by color 3. Thus, $\chi(C_{2n+1}) = 3$.

It is not hard to prove that G is 2-colorable if and only if it does not contain a cycle of odd length. Graphs that are 2-colorable are usually called *bipartite*, since they split into two parts, with all the edges going from one part to the other. The easy characterization ends here, and no simple criterion equivalent to k -colorability is available for $k \geq 3$. This is related to the fact that for each fixed $k \geq 3$ the computational problem of deciding whether a given graph is k -colorable is NP-hard, a notion discussed in COMPUTATIONAL COMPLEXITY [IV.20].

Coloring is one of the most fundamental notions of graph theory, as a huge array of problems in this field and in related areas like computer science and operations research can be formulated in terms of graph coloring. Finding an optimal coloring of a graph is known to be a very hard task, both theoretically and practically.

There are two simple yet fundamental lower bounds on the chromatic number. First, as every color class in a proper coloring of a graph G forms an independent set, it cannot be bigger than the independence number of G , which is denoted by $\alpha(G)$. Therefore, at least $|V(G)|/\alpha(G)$ colors are necessary. Secondly, if G contains a clique of size k , then k colors are needed to color that clique alone, and thus $\chi(G) \geq k$. This implies that $\chi(G) \geq \omega(G)$, where $\omega(G)$ is the clique number of G .

What about upper bounds on the chromatic number? One of the simplest approaches to coloring a graph is to do it *greedily*: put the vertices in some order and color them one by one, assigning to each one the smallest positive integer that has not already been assigned to one of its neighbors. While the greedy algorithm can sometimes be very inefficient (for example, it can color bipartite graphs in an unbounded number of colors, even though two colors are sufficient), it often works quite well. Observe that when applying the greedy algorithm, a color given to a vertex v is at most one more than the number of the neighbors of v preceding it in the chosen order, and is thus at most $d(v) + 1$, where $d(v)$ is the degree of v in G . It follows that, if $\Delta(G)$ is the maximum degree of G , then the greedy algorithm uses at most $\Delta(G) + 1$ colors. Therefore $\chi(G) \leq \Delta(G) + 1$. This bound is tight for complete graphs and

odd cycles, and, as shown by Brooks in 1941, those are the only cases: if G is a graph of maximum degree Δ , then $\chi(G) \leq \Delta$ unless G contains a clique $K_{\Delta+1}$, or $\Delta = 2$ and G contains an odd cycle.

It is also possible to color the *edges* of a graph, rather than the vertices. In this case a proper coloring is defined to be one where no two edges that meet at a vertex are given the same color. The *chromatic index* of G , denoted by $\chi'(G)$, is the minimum k for which G admits a proper edge-coloring with k colors. For example, if G is the complete graph K_{2n} , then $\chi'(G) = 2n - 1$. This turns out to be equivalent to the fact that it is possible to organize a round-robin tournament with $2n$ teams and fit it into $2n - 1$ rounds: just ask the manager of a soccer league. It is also not hard to show that $\chi'(K_{2n-1}) = 2n - 1$. Since in any proper edge-coloring of G all edges of G that are incident to a vertex v get distinct colors, the chromatic index is obviously at least as big as the maximum degree. Equality holds for bipartite graphs, as proved by König in 1931, which implies the existence of a complete timetable using d periods in the problem of teachers and classes discussed in the introduction.

Remarkably, this trivial lower bound of $\chi'(G) \geq \Delta(G)$ is very close to the true behavior of $\chi'(G)$. A fundamental theorem of Vizing from 1964 states that $\chi'(G)$ is always equal either to the maximum degree $\Delta(G)$ or to $\Delta(G) + 1$. Thus, the chromatic index of G is much easier to approximate than its chromatic number.

2.1.2 Excluded Subgraphs

If a graph G has n vertices and contains no triangle (that is, three vertices all joined to each other) then how many edges can it contain? If n is even, then you can split the vertex set into two equal parts A and B of size $n/2$ and join every vertex in A to every vertex in B . The resulting graph G contains no triangles and has $n^2/4$ edges. Moreover, adding another edge will automatically create a triangle (in fact, several triangles). But is this the densest possible triangle-free graph? A hundred years ago the answer was shown to be yes by Mantel. (A similar theorem holds when n is odd, but now A and B must have nearly equal sizes $(n + 1)/2$ and $(n - 1)/2$.)

Let us look at a more general problem, where the role of the triangle is played by an arbitrary graph. More precisely, let H be any graph, with m vertices, say, and when $n \geq m$ let us define $\text{ex}(n, H)$ to be the maximum possible number of edges in a graph with n vertices

that does not contain H as a subgraph. (The notation “ex” stands for “exclude.”) The function $\text{ex}(n, H)$ is usually called the Turán number of H , for reasons that will become clear, and finding good approximations for it has been a central problem in extremal graph theory.

What kind of examples of graphs that do not contain H can we think of? One observation that gets us started is that if H has chromatic number r , then it cannot be a subgraph of a graph G with chromatic number less than r . (Why not? Because a proper $(r - 1)$ -coloring of G provides us with a proper $(r - 1)$ -coloring of any subgraph of G .) So a promising approach is to look for a graph G with n vertices, chromatic number $r - 1$, and as many edges as possible. This is easy to find. Our constraint is that the vertices can be partitioned into $r - 1$ independent sets. Once we have done that, we may as well include all edges between those sets. The result is a *complete $(r - 1)$ -partite* graph. A routine calculation shows that in order to maximize the number of edges, one should partition into sets that have sizes as nearly equal as possible. (For example, if $n = 10$ and $r = 4$, then we would partition into three sets of sizes 3, 3, and 4.)

The graph that satisfies this condition is called the *Turán graph* $T_{r-1}(n)$ and its number of edges is denoted by $t_{r-1}(n)$. We have just argued that $\text{ex}(n, H) \geq t_{r-1}(n)$, which can be shown to be at least as big as $(1 - 1/(r - 1))\binom{n}{2}$.

Turán’s contribution to this area was to give an exact solution, in 1941, for the most important case, when H is the complete graph K_r on r vertices. He proved that $\text{ex}(n, K_r)$ is not just at least $t_{r-1}(n)$, but is actually equal to $t_{r-1}(n)$. Moreover, the only K_r -free graph with n vertices and $\text{ex}(n, K_r)$ edges is the Turán graph $T_{r-1}(n)$. Turán’s paper is generally considered the starting point of extremal graph theory.

Later, Erdős, Stone, and Simonovits extended Turán’s theorem by proving that the above simple lower bound for $\text{ex}(n, H)$ is *asymptotically* tight for any fixed H with chromatic number at least 3. That is, if r is the chromatic number of H , then the ratio of $\text{ex}(n, H)$ to $t_{r-1}(n)$ tends to 1 as n tends to infinity.

Thus, the function $\text{ex}(n, H)$ is well-understood for all nonbipartite graphs. Bipartite graphs are rather different, because their Turán numbers are much smaller: if H is bipartite, then $\text{ex}(n, H)/n^2$ tends to zero. Determining the asymptotics of $\text{ex}(n, H)$ in this case remains a challenging open problem with many unsettled questions. Indeed, the full story is unknown even for the very simple case when H is a cycle. Partial results

obtained so far use a variety of techniques from different fields, including probability theory, number theory, and algebraic geometry.

2.1.3 Matchings and Cycles

Let G be a graph. A *matching* in G is a collection of edges in G of which no two share a vertex. A matching M in G is called *perfect* if every vertex belongs to one of the edges in M . (The idea is that the edges determine a “match” for each vertex: the match for x is the vertex y for which xy is an edge of M .) Of course, for G to have a perfect matching it must have an even number of vertices.

One of the best-known theorems in graph theory is Hall’s theorem, which provides a necessary and sufficient condition for the existence of a perfect matching in a bipartite graph. What kind of condition can this be? It is very easy to write down a trivial *necessary* condition, as follows. Let G be a bipartite graph with vertex sets A and B of equal size. (If they do not have equal size, then clearly there is no perfect matching.) Given any subset S of A , let $N(S)$ denote the set of all vertices in B that are joined to at least one vertex in S . If there is to be a matching, then it must be possible to assign to each vertex in S a distinct “match,” so obviously $N(S)$ must have at least as many elements as S . Hall’s theorem, proved in 1935, asserts that, remarkably, this obvious necessary condition is also sufficient. That is, if $N(S)$ is always at least as big as S , then there will be a perfect matching. More generally, if A is smaller than B , then the same condition guarantees that one can find a matching that includes every vertex in A (but leaves some vertices in B unmatched).

There is a useful reformulation of Hall’s theorem in terms of set systems. Let S_1, S_2, \dots, S_n be a collection of sets, and suppose that we would like to find a system of *distinct representatives*: that is, a sequence x_1, x_2, \dots, x_n such that x_i is an element of S_i and no two of the x_i are the same. Obviously this cannot be done if the union of some k of the sets S_i has size less than k . Again, this obvious necessary condition is sufficient. It is not hard to show that this assertion is equivalent to Hall’s theorem: let S be the union of the S_i and define a bipartite graph with vertex sets $\{1, 2, \dots, n\}$ and S , joining i to x if and only if $x \in S_i$. Then a matching that includes all of the set $\{1, 2, \dots, n\}$ picks out a system of distinct representatives: x_i is the element of S that is matched with i .

Hall’s theorem can be applied to solve the problem of finding a system of representatives for the right and

left cosets of a subgroup H , mentioned in section 1.1. Define a bipartite graph F , whose two sides (of size n/k each) are the left and right cosets of H . A left coset g_1H is connected by an edge of F to a right coset Hg_2 if they share a common element. It is not difficult to show that F satisfies the Hall condition, and hence it has a perfect matching M . Choosing for each edge (g_iH, Hg_j) of M a common element of g_iH and Hg_j , we obtain the required family of representatives.

There is also a necessary and sufficient condition for the existence of a perfect matching in a general (not-necessarily-bipartite) graph G . This is a theorem of Tutte, which we shall not state here.

Recall that C_k denotes a cycle of length k . A cycle is a very basic graph structure, and, as one might expect, there are many extremal results concerning cycles.

Suppose that G is a connected graph with no cycles. If you pick a vertex and look at its neighbors and then the neighbors of its neighbors, and so on, you will see that it has a tree-like structure. Indeed, such graphs are called *trees*. An easy exercise shows that any tree with n vertices has exactly $n - 1$ edges. It follows that every graph G on n vertices with at least n edges has a cycle. If you want to guarantee that this cycle has certain extra properties, then you may need more edges. For example, the theorem of Mantel mentioned earlier implies that a graph G with n vertices and more than $n^2/4$ edges contains a triangle $C_3 = K_3$. One can also prove that a graph $G = (V, E)$ with $|E| > \frac{1}{2}k(|V| - 1)$ has a cycle of length longer than k (and this is in fact a sharp result).

A *Hamilton cycle* in a graph G is a cycle that visits every vertex of G . This term originated in a game, invented by HAMILTON [VI.37] in 1857, the objective of which was to complete a Hamilton cycle in the graph of the dodecahedron. A graph containing a Hamilton cycle is called *Hamiltonian*. This concept is strongly related to the well-known TRAVELING SALESMAN PROBLEM [VII.5 §2]: you are given a graph with positive weights assigned to the edges, and you must find a Hamilton cycle for which the sum of the weights of its edges is minimized. There are many sufficient criteria for a graph to be Hamiltonian, quite a few of which are based on the sequence of degrees. For example, Dirac proved in 1952 that a graph on $n \geq 3$ vertices all of whose degrees are at least $n/2$ is Hamiltonian.

2.2 Ramsey Theory

Ramsey theory is a systematic study of the following general phenomenon. Surprisingly often, a large struc-

ture of a certain kind has to contain a fairly large highly organized substructure, even if the structure itself is completely arbitrary and apparently chaotic. As succinctly put by the mathematician T. S. Motzkin, "Complete disorder is impossible." One might expect that the simple and very general form of this paradigm ensures that it has many diverse manifestations in different mathematical areas, and this is indeed the case. (One should, however, bear in mind that some natural statements of this kind are false for nonobvious reasons.)

A very simple statement, which can be regarded as a basic prototype for what follows, is the *pigeonhole principle*. This states that if a set X of n objects is colored with s colors, then there must be a subset of X of size at least n/s that uses just one color. Such a subset is called *monochromatic*.

The situation becomes more interesting if the set X has some additional structure. It then becomes natural to ask for a monochromatic subset that keeps some of the structure of X . However, it also becomes much less obvious whether such a subset exists. Ramsey theory consists of problems and theorems of this general kind. Although several Ramsey-type theorems had appeared before, Ramsey theory is traditionally regarded as having started with *Ramsey's theorem*, proved in 1930. Ramsey took as his set X the set of all the edges in a complete graph, and the monochromatic subset he obtained consisted of all the edges of some complete subgraph. A precise statement of his theorem is as follows. Let k and l be integers greater than 1. Then there exists an integer n such that, however you color the edges of the complete graph with n vertices, using the two colors red and blue, there will either be k vertices such that all edges between them are red or l vertices such that all edges between them are blue. That is, a sufficiently large complete graph colored with two colors contains a largish complete subgraph that is monochromatic. Let $R(k, l)$ denote the minimum number n with this property. In this language, the observation of Szalai, mentioned in the introduction, is that $R(4, 4) \leq 20$ (in fact, $R(4, 4) = 18$). Actually, Ramsey's theorem was more general, in that he allowed any number of colors, and the objects colored could be r -tuples of elements rather than just pairs, as one has when coloring graphs. The exact computation of small Ramsey numbers turns out to be a notoriously difficult task: even the value of $R(5, 5)$ is unknown at present.

The second cornerstone of Ramsey theory was laid by Erdős and Szekeres, who in 1935 wrote a paper

PUP: Tim says in answer to the proofreaders's comment here, "It's the contrapositive - no contradiction." OK?

containing several important Ramsey-type results. In particular, they proved the recursion $R(k, l) \leq R(k-1, l) + R(k, l-1)$. Combined with the easy boundary conditions $R(2, l) = l$, $R(k, 2) = k$, the recursion leads to the estimate $R(k, l) \leq \binom{k+l-2}{k-1}$. In particular, for the so-called diagonal case $k = l$ we obtain $R(k, k) < 4^k$. Remarkably, no improvement in the exponent of the latter estimate has been found so far. That is, nobody has found an upper bound of the form C^k for some $C < 4$. The best lower bound known, which we shall discuss in section 3.2, is roughly $R(k, k) \geq 2^{k/2}$, so there is a rather substantial gap.

Another Ramsey-type statement, proved by Erdős and Szekeres, is of a geometric nature. They showed that for every $n \geq 3$ there exists a positive integer N such that, given any configuration of N points in the plane in general position (i.e., no three of them are on a line), there are n that form a convex n -gon. (It is instructive to prove that if $n = 4$ then N can be taken to be 5.) There are several proofs of this theorem, some using the general Ramsey theorem. It is conjectured that the smallest value of N that will do in order to ensure a convex n -gon is $2^{n-2} + 1$.

The classic Erdős-Szekeres paper also contains the following Ramsey-type result: any sequence of $n^2 + 1$ distinct numbers contains a monotone (increasing or decreasing) subsequence of length $n + 1$.

This provides a quick lower bound of \sqrt{n} for a well-known problem of Ulam, asking for the typical length of a longest increasing subsequence of a random sequence of length n . A detailed description of the distribution of this length has recently been given by Baik, Deift, and Johansson.

In 1927 van der Waerden proved what became known as van der Waerden's theorem: for all positive integers k and r there exists an integer W such that for every coloring of the set of integers $\{1, \dots, W(k, r)\}$ using r colors, one of the colors contains an arithmetic progression of length k . The minimum W for which this is true is denoted by $W(k, r)$. Van der Waerden's bounds for $W(k, r)$ are enormous: they grow like an Ackermann-type function. A new proof of his theorem was found by Shelah in 1987, and yet another proof was given by Gowers in 2000, while he was studying the (much deeper) "density version" of the theorem, which will be described in section 2.4. These recent proofs provided improved upper bounds for $W(k, r)$, but the best-known lower bound for this number, which is only exponential in k for each fixed r , is much smaller.

Even before van der Waerden, Schur proved in 1916 that for any positive integer r there exists an integer $S(r)$ such that for every r -coloring of $\{1, \dots, S(r)\}$ one of the colors contains a solution of the equation $x + y = z$. The proof can be derived rather easily from the general Ramsey theorem. Schur applied this statement to prove the following result, mentioned in section 1.1: for every k and all sufficiently large primes p , the equation $a^k + b^k = c^k$ has a nontrivial solution in the integers modulo p . To prove this result, assume that $p \geq S(k)$ and consider the FIELD [I.3 §2.2] \mathbb{Z}_p of integers mod p . The nonzero elements of \mathbb{Z}_p form a GROUP [I.3 §2.1] under multiplication. Let H be the subgroup of this group consisting of all k th powers: that is, $H = \{x^k : x \in \mathbb{Z}_p^*\}$. It is not hard to show that the index r of H is the highest common factor of k and $p-1$, and in particular is at most k . The partition of \mathbb{Z}_p^* into the cosets of H can be thought of as an r -coloring of \mathbb{Z}_p^* . By Schur's theorem there exist $x, y, z \in \{1, \dots, p-1\}$ that all have the same color—that is, they all belong to the same coset of H . In other words, there exists a residue $d \in \mathbb{Z}_p^*$ such that $x = da^k$, $y = db^k$, $z = dc^k$, and $da^k + db^k = dc^k$ modulo p . The desired result follows if we multiply both sides by d^{-1} .

Many additional Ramsey-type results can be found in Graham, Rothschild, and Spencer (1990) or in Graham, Grötschel, and Lovász (1995, chapter 25).

2.3 Extremal Theory of Set Systems

Graphs are one of the fundamental structures studied by combinatorialists, but there are others too. An important branch of the subject is the study of *set systems*. Most often, these are simply collections of subsets of some n -element set. For example, the collection of all subsets of the set $\{1, 2, \dots, n\}$ of size at most $n/3$ is a good example of a set system. An extremal problem in this area is any problem where the aim is to determine, or estimate, the maximum number of sets there can be in a set system that satisfies certain conditions. For example, one of the first results in the area was proved by Sperner in 1928. He looked at the following question: how large a collection of subsets can one choose from an n -element set in such a way that no set from the collection is a subset of any other? A simple example of a set system satisfying this condition is the collection of all sets of size r , for some r . From this it immediately follows that we can obtain a collection as large as the largest binomial coefficient, which is $\binom{n}{n/2}$ if n is even and $\binom{n}{(n+1)/2}$ if n is odd.

Sperner showed that this is indeed the maximum possible size of such a collection. This result supplies a quick solution to the real analogue of the problem of Littlewood and Offord described in section 1.1. Suppose that x_1, x_2, \dots, x_n are n not-necessarily-distinct real numbers, each of modulus at least 1. A first observation is that we may assume that all the x_i are positive, since if we replace a negative x_i by $-x_i$ (which is positive), then we end up with exactly the same set of sums, but shifted by $-x_i$. (To see this, compare a sum that used to involve x_i with the corresponding sum that does not involve $-x_i$, and vice versa.) But now, if A is a proper subset of B , then some x_i belongs to B and not to A , so

$$\sum_{i \in B} x_i - \sum_{i \in A} x_i \geq x_i \geq 1.$$

Therefore, the total number of subset sums you can find with any two differing by less than 1 is at most $\binom{n}{\lfloor n/2 \rfloor}$, by Sperner's theorem.

A set system is called an *intersecting family* if any two sets in the system intersect. Since a set and its complement cannot both belong to an intersecting family of subsets of $\{1, 2, \dots, n\}$, we see immediately that such a family can have size at most 2^{n-1} . Moreover, this bound is achieved by, for example, the collection of all sets that contain the element 1. But what happens if we fix a k and assume in addition that all our sets have size k ? We may assume that $n \geq 2k$, as otherwise the solution is trivial. Erdős, Ko, and Rado proved that the maximum is $\binom{n-1}{k-1}$. Here is a beautiful proof discovered later by Katona. Suppose you arrange the elements randomly around a circle. Then there are n ways of choosing k elements that are consecutive in this arrangement, and it is quite easy to convince yourself that at most k of these can intersect (if $n \geq 2k$). So out of these n sets of size k , only k of them can belong to any given intersecting family. Now it is also easy to show that every set has an equal chance of being one of these n sets, and this proves (by a simple double-counting argument) that the largest possible proportion of sets in the family is k/n . Therefore, the family itself has size at most $(k/n) \binom{n}{k}$, which equals $\binom{n-1}{k-1}$. The original proof of Erdős, Ko, and Rado is more complicated than this, but it is important because it introduced a technique known as *compression*, which was used to solve many other extremal problems.

Let $n > 2k$ be two positive integers. Suppose that you wish to color all subsets of the set $\{1, 2, \dots, n\}$ of size k in such a way that any two sets with the same color intersect each other. What is the smallest number of

colors you can use? It is not difficult to see that $n - 2k + 2$ colors suffice. Indeed, one color class can be the family of all subsets of $\{1, 2, \dots, 2k - 1\}$, which is clearly an intersecting family. And then, for each i such that $2k \leq i \leq n$, you can take the family of all subsets whose largest element is i . There are $n - 2k + 1$ such families, and any set of size k belongs either to one of them or to the first family. Therefore, $n - 2k + 2$ colors are enough.

Kneser conjectured in 1955 that this bound was tight: in other words, that if you have fewer than $n - 2k + 2$ colors then you will have to give the same color to some pair of disjoint sets. This conjecture was proved by Lovász in 1978. His proof is topological, and uses the Borsuk-Ulam theorem. Several simpler proofs have been found since, but they are all based on the topological idea in the first proof. Since Lovász's breakthrough, topological arguments have become an important part of the armory of researchers in combinatorics.

2.4 Combinatorial Number Theory

Number theory is one of the oldest branches of mathematics. At its core are problems about integers, but a sophisticated array of techniques has been developed to deal with those problems, and these techniques have often themselves been the basis for further study (see, for example, ALGEBRAIC NUMBERS [IV.1], ANALYTIC NUMBER THEORY [IV.2], and ARITHMETIC GEOMETRY [IV.5]). However, some problems in number theory have yielded to the methods of combinatorics. Some of these problems are extremal problems with a combinatorial flavor, while others are classical problems in number theory where the existence of a combinatorial solution has been quite surprising. We describe below a few examples. Many more can be found in chapter 20 of Graham, Grötschel, and Lovász (1995), in Nathanson (1996), and in Tao and Vu (2006).

A simple but important notion in the area is that of a *sumset*. If A and B are two sets of integers, or more generally are two subsets of an ABELIAN GROUP [I.3 §2.1], then the sumset $A + B$ is defined to be $\{a + b : a \in A, b \in B\}$. For instance, if $A = \{1, 3\}$ and $B = \{5, 6, 12\}$, then $A + B = \{6, 7, 8, 9, 13, 15\}$. There are many results relating the size and structure of $A + B$ to those of A and B . For example, the *Cauchy-Davenport theorem*, which has numerous applications in additive number theory, is the statement that if p is a prime, and A, B are two nonempty subsets of \mathbb{Z}_p , then the size of $A + B$ is at least the minimum of p and $|A| + |B| - 1$. (Equality occurs if A and B are arithmetic progressions with the

same common difference.) CAUCHY [VI.29] proved this theorem in 1813, and applied it to give a new proof of a lemma that LAGRANGE [VI.22] had proved as part of his well-known 1770 paper that shows that every positive integer is a sum of four squares. Davenport formulated the theorem as a discrete analogue of a related conjecture of Khinchin about densities of sums of sequences of integers. The proofs given by Cauchy and by Davenport are combinatorial, but there is also a more recent algebraic proof, based on some properties of roots of polynomials. The advantage of the latter is that it provides many variants that do not seem to follow from the combinatorial approach. For example, let us define $A \oplus B$ to be the set of all $a+b$ such that $a \in A$, $b \in B$, and $a \neq b$. Then the smallest possible size of $A \oplus B$, given the sizes of A and B , is the minimum of p and $|A| + |B| - 2$. Further extensions can be found in Nathanson (1996) and in Tao and Vu (2006).

The theorem of van der Waerden mentioned in section 2.2 implies that, however you color the positive integers with some finite number r of colors, there must be some color that contains arithmetic progressions of every length. Erdős and Turán conjectured in 1936 that this always holds for the “most popular” color class. More precisely, they conjectured that for any positive integer k and for any real number $\epsilon > 0$, there is a positive integer n_0 such that if $n > n_0$, any set of at least ϵn positive integers between 1 and n contains a k -term arithmetic progression. (Setting $\epsilon = r^{-1}$ one can easily deduce van der Waerden’s theorem from this.) After several partial results, this conjecture was proved by Szemerédi in 1975. His deep proof is combinatorial, and applies techniques from Ramsey theory and extremal graph theory. Furstenberg gave another proof in 1977, based on techniques of ERGODIC THEORY [V.11]. In 2000 Gowers gave a new proof, combining combinatorial arguments with tools from analytic number theory. This proof supplied a much better quantitative estimate. A related very recent spectacular result of Green and Tao asserts that there are arbitrarily long arithmetic progressions of prime numbers. Their proof combines number-theoretic techniques with the ergodic theory approach. Erdős conjectured that any infinite sequence n_i for which the sum $\sum_i (1/n_i)$ diverges contains arbitrarily long arithmetic progressions. This conjecture would imply the theorem of Green and Tao.

2.5 Discrete Geometry

Let P be a set of points and let L be a set of lines in the plane. Let us define an *incidence* to be a pair (p, ℓ) , where p is a point in P , ℓ is a line in L , and the point p lies on the line ℓ . Suppose that P contains m distinct points and L contains n distinct lines. How many incidences can there be? This is a geometrical problem, but again it has a strong flavor of extremal combinatorics. As such, it is typical of the area known as *discrete* (or *combinatorial*) geometry.

Let us write $I(m, n)$ for the maximum number of incidences there can be between m points and n lines. Szemerédi and Trotter determined the asymptotic behavior of this quantity, up to a constant factor, for all possible values of m and n . There are two absolute positive constants c_1, c_2 such that, for all m, n ,

$$c_1(m^{2/3}n^{2/3} + m + n) \leq I(m, n) \leq c_2(m^{2/3}n^{2/3} + m + n).$$

If $m > n^2$ or $n > m^2$ then one can establish the lower bound by taking all m points on a single line, or all n lines through a single point, respectively. In the harder cases when m and n are closer to each other, one can prove it by letting P contain all the points of a $\lfloor \sqrt{m} \rfloor$ by $\lfloor \sqrt{m} \rfloor$ grid, and by taking the n most “popular” lines: that is, the n lines that contain the most points of P . Establishing the upper bound is more difficult. The most elegant proof of it is due to Székely, and is based on the fact that, however you draw a graph with m vertices and more than $4m$ edges, you must have many pairs of edges that cross each other. (This is a rather simple consequence of the famous Euler formula connecting the numbers of vertices, edges, and regions in any drawing of a planar graph.) To bound the number of incidences between a set of points P and a set of lines L in the plane, one considers the graph whose vertices are the points P , and whose edges are all segments between consecutive points along a line in L . The desired bound is obtained by observing that the number of crossings in this graph does not exceed the number of pairs of lines in L , and yet should be large if there are many incidences.

Similar ideas can be used to give a partial answer to the following question: if you take n points in the plane, how many pairs (x, y) of these points can there be with the distance from x to y equal to 1? It is not surprising that the two problems are related: the number of such pairs is the number of incidences between the given n points and the n unit circles that are centered at these

PUP: I can confirm that it's OK that the parenthetical term on the LHS of line 1 is the same as the parenthetical term on the RHS of line 2 here.

points. Here, however, there is a large gap between the best known upper bound, which is $cn^{4/3}$ for some absolute constant c , and the best known lower bound, which is only $n^{1+c'/\log \log n}$ for some constant $c' > 0$.

A fundamental theorem of Helly asserts that if you have a finite family of at least $d + 1$ convex sets in \mathbb{R}^d , and if any $d + 1$ of them have a point in common, then all sets in the family have a common point. Now let us start with a weaker assumption: given any p of the sets, some $d + 1$ of those p sets have a point in common. (Here p is some integer greater than $d + 1$.) Can one then find a set X of at most C points such that each set in \mathcal{F} contains a point in X , with C a constant that depends on p but not on the number of convex sets in the family? This question was raised by Hadwiger and Debrunner in 1957 and solved by Kleitman and Alon in 1992. The proof combines a “fractional version” of Helly’s theorem with the duality of linear programming (see OPTIMIZATION AND LAGRANGE MULTIPLIERS [III.66]) and various additional geometric results. Unfortunately, it gives a very poor estimate for C : even in two dimensions and with $p = 4$ it is not known what the best possible value of C is.

This is just a small sample of problems and results in discrete geometry. Such results have been applied extensively in computational geometry and in combinatorial optimization in recent decades. Two good books on the subject are Pach and Agarwal (1995) and Matoušek (2002).

2.6 Tools

Many of the basic results in extremal combinatorics were obtained mainly by ingenuity and detailed reasoning. However, the subject has grown out of this early stage: several deep tools have been developed that have been essential to much of the recent progress in the area. In this subsection, we include a very brief description of some of these tools.

Szemerédi’s regularity lemma is a result in graph theory that has numerous applications in various areas, including combinatorial number theory, computational complexity, and, mainly, extremal graph theory. The precise statement of the lemma, which can be found, for example, in Bollobás (1978), is somewhat technical. The rough statement is that the vertex set of any large graph can be partitioned into a constant number of pieces of nearly equal size, so that the bipartite graphs between most pairs of pieces behave like random bipartite graphs. The strength of this lemma is that it applies

to any graph, providing a rough approximation of its structure that enables one to extract a lot of information about it. A typical application is that a graph with “few” triangles can be “well-approximated” by a graph with no triangles. More precisely, for any $\epsilon > 0$ there exists $\delta > 0$ such that if G is a graph with n vertices and at most δn^3 triangles, then one can remove at most ϵn^2 edges from G and make it triangle free. This innocent-looking statement turns out to imply the case $k = 3$ of Szemerédi’s theorem that was mentioned earlier.

Tools from linear and multilinear algebra play an essential role in extremal combinatorics. The most fruitful technique of this kind, which is possibly also the simplest, is the so-called *dimension argument*. In its simplest form, the method can be described as follows. In order to bound the cardinality of a discrete structure A , one maps its elements to distinct vectors in a VECTOR SPACE [I.3 §2.3], and proves that those vectors are linearly independent. It then follows that the size of A is at most the dimension of the vector space in question. An early application of this argument was found by Larman, Rogers, and Seidel in 1977. They wanted to know how many points it was possible to find in \mathbb{R}^n that determine at most two distinct differences. An example of such a system is the set of all points whose coordinates consist of $n - 2$ 0s and two 1s. Notice, however, that these points all lie in the hyperplane of points whose coordinates add up to 2. So this actually provides us with an example in \mathbb{R}^{n-1} . Therefore, we have a simple lower bound of $n(n + 1)/2$. Larman, Rogers, and Seidel matched this with an upper bound of $(n + 1)(n + 4)/2$. They did this by associating with each point of such a set a polynomial in n variables, and by showing that these polynomials are linearly independent and all lie in a space of dimension $(n + 1)(n + 4)/2$. This has been improved by Blokhuis to $(n + 1)(n + 2)/2$. He did this by finding $n + 1$ further polynomials that lie in the same space in such a way that the augmented set of polynomials is still linearly independent. More applications of the dimension argument can be found in Graham, Grötschel, and Lovász (1995, chapter 31).

Spectral techniques, that is, an analysis of EIGEN-VECTORS AND EIGENVALUES [I.3 §4.3], have been used extensively in graph theory. The link comes through the notion of an *adjacency matrix* of a graph G . This is defined to be the matrix A with entries $a_{u,v}$ for each pair of (not-necessarily-distinct) vertices u and v , where $a_{u,v} = 1$ if u and v are joined by an edge, and $a_{u,v} = 0$ otherwise. This matrix is symmetric, and therefore, by standard results in linear algebra, it has

PUP: Tim says that the singular is correct here.

real eigenvalues and an ORTHONORMAL BASIS [III.37] of eigenvectors. It turns out that there is a tight relationship between the eigenvalues of the adjacency matrix A and several structural properties of the graph G , and these properties can often be useful in the study of various extremal problems. Of particular interest is the second largest eigenvalue of a regular graph. Suppose that every vertex of a graph G has degree d . Then the vector for which every entry is 1 is easily seen to be an eigenvector with eigenvalue d , and this is the largest eigenvalue. If all other eigenvalues have modulus much smaller than d , then it turns out that G behaves in many ways like a random d -regular graph. In particular, the number of edges inside any set of k of the vertices is roughly the same (provided k is not too small) as one would expect with a random graph. It follows easily that any set of vertices that is not too big has many neighbors among the vertices outside that set. Graphs with the latter property are called EXPANDERS [III.24] and have numerous applications in theoretical computer science. Constructing such graphs explicitly is not an easy matter and was at one time a major open problem. Now, however, several constructions are known, based on algebraic tools. See chapter 9 of Alon and Spencer (2000), and its references, for more details.

The application of topological methods in the study of combinatorial objects such as partially ordered sets, graphs, and set systems has already become part of the mathematical machinery commonly used in combinatorics. An early example is Lovász's proof of Kneser's conjecture, mentioned in section 2.3. Another example is a result of which the following is a representative special case. Suppose you have a piece of string with 10 red beads, 15 blue beads, and 20 yellow beads on it. Then, no matter what order the beads come in, you can cut the string in at most 12 places and place the resulting segments of beaded string into five piles, each of which contains two red beads, three blue beads, and four yellow beads. The number 12 is obtained by multiplying 4, the number of piles minus 1, by 3, the number of colors. The general case of this result was proved by Alon using a generalization of Borsuk's theorem. Many additional examples of topological proofs appear in Graham, Grötschel, and Lovász (1995, chapter 34).

3 Probabilistic Combinatorics

A wonderful development took place in twentieth-century mathematics when it was realized that it is sometimes possible to use probabilistic reasoning to prove

mathematical statements that do not have an obvious probabilistic nature. For example, in the first half of the century, Paley, Zygmund, Erdős, Turán, Shannon, and others used probabilistic reasoning to obtain striking results in analysis, number theory, combinatorics, and information theory. It soon became clear that the so-called *probabilistic method* is a very powerful tool for proving results in discrete mathematics. The early results combined combinatorial arguments with fairly elementary probabilistic techniques, but in recent years the method has been greatly developed, and now it often requires one to apply much more sophisticated techniques. A recent text dealing with the subject is Alon and Spencer (2000).

The applications of probabilistic techniques in discrete mathematics were initiated by Paul Erdős, who contributed to the development of the method more than anyone else. One can classify them into three groups.

The first deals with the study of certain classes of random combinatorial objects, like random graphs or random matrices. The results here are essentially results in probability theory, although most of them are motivated by problems in combinatorics. A typical problem is the following: if we pick a graph "at random," what is the probability that it contains a Hamilton cycle?

The second group consists of applications of the following idea. Suppose you want to prove that a combinatorial structure exists with certain properties. Then one possible method is to choose a structure randomly (from a probability distribution that you are free to specify) and estimate the probability that it has the properties you want. If you can show that this probability is greater than 0, then such a structure exists. Surprisingly often it is much easier to prove this than it is to give an example of a structure that works. For instance, is there a graph with large girth (meaning it has no short cycles) and large chromatic number? Even if "large" means "at least 7," it is very hard to come up with an example of such a graph. But their existence is a fairly easy consequence of the probabilistic method.

PUP: Tim thinks the text is fine as it is. OK?

The third group of applications is perhaps the most striking of all. There are many examples of statements that appear to be completely deterministic (even when one is used to the idea of using probability to give existence proofs) but that nevertheless yield to probabilistic reasoning. In the remainder of this section we shall briefly describe some typical examples of each of these three kinds of application.

3.1 Random Structures

The systematic study of random graphs was initiated by Erdős and Rényi in 1960. The most common way of defining a random graph is to fix a probability p and then to join each pair of vertices with an edge with probability p , with all the choices made independently. The resulting graph is denoted $G(n, p)$. (Formally speaking, $G(n, p)$ is not a graph but a probability distribution, but one often talks about it as though it is a graph that has been produced in a random way.) Given any property, such as “contains no triangles,” we can study the probability that $G(n, p)$ has that property.

A striking discovery of Erdős and Rényi was that many properties of graphs “emerge very suddenly.” Some examples are “contains a Hamilton cycle,” “is not planar,” and “is connected.” These properties are all *monotone*, which means that if a graph G has the property and you add an edge to G , then the resulting graph still has the property. Let us take one of these properties and define $f(p)$ to be the probability that the random graph $G(n, p)$ has it. Because the property is monotone, $f(p)$ increases as p increases. What Erdős and Rényi discovered was that almost all of this increase happens in a very short time. That is, $f(p)$ is almost 0 for small p and then suddenly changes very rapidly and becomes almost 1.

Perhaps the most famous and illustrative example of this swift change is the sudden appearance of the so-called *giant component*. Let us look at $G(n, p)$ when p has the form c/n . If $c < 1$, then with high probability all the connected components of $G(n, p)$ have size at most logarithmic in n . However, if $c > 1$, then $G(n, p)$ almost certainly has one component of size linear in n (the giant component), while all the rest have logarithmic size. This is related to the phenomenon of *phase transitions* in mathematical physics, which are discussed in PROBABILISTIC MODELS OF CRITICAL PHENOMENA [IV.25]. A result of Friedgut shows that the phase transition for a graph property that is “global,” in a sense that can be made precise, is sharper than the one for a “local” property.

Another interesting early discovery in the study of random graphs was that many of the basic parameters of graphs are highly “concentrated.” A striking example that illustrates what this means is the fact that, for any fixed value of p and for most values of n , almost all graphs $G(n, p)$ have the same clique number. That is, there exists some r (depending on p and n) such that with high probability, when n is large, the clique

number of $G(n, p)$ is equal to r . Such a result cannot hold for all n , for continuity reasons, but in the exceptional cases there is still some r such that the clique number is almost certainly equal either to r or to $r + 1$. In both cases, r is roughly $2 \log n / \log(1/p)$. The proof of this result is based on the so-called *second moment method*: one estimates the expectation and the variance of the number of cliques of a given size contained in $G(n, p)$, and applies well-known inequalities of Markov and CHEBYSHEV [VI.45].

The chromatic number of the random graph $G(n, p)$ is also highly concentrated. Its typical behavior for values of p that are bounded away from 0 was determined by Bollobás. A more general result, in which p is allowed to tend to 0 as $n \rightarrow \infty$, was proved by Shamir, Spencer, Łuczak, Alon, and Krivelevich. In particular, it can be shown that for every $\alpha < \frac{1}{2}$ and every integer-valued function $r(n) < n^\alpha$, there exists a function $p(n)$ such that the chromatic number of $G(n, p(n))$ is precisely $r(n)$ almost surely. However, determining the precise degree of concentration of the chromatic number of $G(n, p)$, even in the most basic and important case $p = \frac{1}{2}$ (in which all labeled graphs on n vertices occur with equal probability), remains an intriguing open problem.

Many additional results on random graphs can be found in Janson, Łuczak, and Ruciński (2000).

3.2 Probabilistic Constructions

One of the first applications of the probabilistic method in combinatorics was a lower bound given by Erdős for the Ramsey number $R(k, k)$, which was defined in section 2.2. He proved that if

$$\binom{n}{k} 2^{1-\binom{k}{2}} < 1,$$

then $R(k, k) > n$. That is, there is a red/blue coloring of the edges of the complete graph on n vertices such that no clique of size k is completely red or completely blue. Notice that the number $n = \lfloor 2^{k/2} \rfloor$ satisfies the above inequality for all $k \geq 3$, so Erdős’s result gives an exponential lower bound for $R(k, k)$. The proof is simple: if you color the edges randomly and independently, then the probability that any fixed set of k vertices has all its edges of the same color is twice $2^{-\binom{k}{2}}$. Thus, the expected number of cliques with this property is

$$\binom{n}{k} 2^{1-\binom{k}{2}}.$$

If this is less than 1, then there must be at least *some* colorings for which there are no cliques with this property, and the result is proved.

Note that this proof is completely nonconstructive, in the sense that it merely proves the existence of such a coloring, but gives no efficient way of actually constructing one.

A similar computation yields a solution for the tournament problem mentioned in section 1.1. If the results of the tournament are random, then the probability, for any particular k teams, that no other team beats them all is $(1 - (1/2^k))^{n-k}$. From this it follows that if

$$\binom{n}{k} \left(1 - \frac{1}{2^k}\right)^{n-k} < 1,$$

then there is a nonzero probability that for every choice of k teams, there is another team that beats them all. In particular, it is possible for this to happen. If n is larger than about $k^2 2^k \log 2$, then the above inequality holds.

Probabilistic constructions have been very powerful in supplying lower bounds for Ramsey numbers. Besides the bound for $R(k, k)$ mentioned above, there is a subtle probabilistic proof, due to Kim, that $R(3, k) \geq ck^2 / \log k$, for some $c > 0$. This is known to be tight up to a constant factor, as proved by Ajtai, Komlós, and Szemerédi, who also used probabilistic methods.

3.3 Proving Deterministic Theorems

Suppose that you color the integers with k colors. Let us call a set S *multicolored* if all k colors appear in S . Straus conjectured that for every k there is an m with the following property: given any set S with m elements, there is a coloring of the integers with k colors such that all translates of S are multicolored. This conjecture was proved by Erdős and Lovász. The proof is probabilistic, and applies a tool called the *Lovász local lemma*, which, unlike many probabilistic techniques, allows one to show that certain events hold with nonzero probability even when this probability is extremely small. The assertion of this lemma, which has numerous additional applications, is, roughly, that for any finite collection of “nearly independent” low-probability events, there is a positive probability that none of the events holds. Note that the statement of Straus’s conjecture has nothing to do with probability, and yet its proof relies on probabilistic arguments.

A graph G is k -colorable, as we have said, if you can properly color its vertices with k colors. Suppose now that instead of trying to use k colors in total, you have

a separate list of k colors for each vertex, and this time you want to find a proper coloring of G where each vertex gets a color from its own list. If you can always do so, no matter what the lists are, then G is called *k-choosable*, and the smallest k for which G is *k-choosable* is called the *choice number* $\text{ch}(G)$. If all the lists are the same, then one obtains a k -coloring, so $\text{ch}(G)$ must be at least as big as $\chi(G)$. One might expect $\text{ch}(G)$ to be equal to $\chi(G)$, since it seems as though using different lists of k colors for different vertices would make it easier to find a proper coloring than using the same k colors for all vertices. However, this turns out to be far from true. It can be proved that for any constant c there is a constant C such that any graph with average degree at least C has choice number at least c . Such a graph might easily be bipartite (and therefore have chromatic number 2), so it follows that $\text{ch}(G)$ can be much bigger than $\chi(G)$. Somewhat surprisingly, the proof of this result is probabilistic.

An interesting application of this fact concerns a graph that arises in Ramsey theory. Its vertices are all the points in the plane, with two vertices joined by an edge if and only if the distance between them is 1. The choice number of this graph is infinite, by the above result, but the chromatic number is known to be between 4 and 7.

A typical problem in Ramsey theory asks for a substructure of some kind that is entirely colored with one color. Its cousin, *discrepancy theory*, merely asks that the numbers of times the colors are used are not too close to each other. Probabilistic arguments have proved extremely useful in numerous problems of this general kind. For example, Erdős and Spencer proved that in any red/blue coloring of the edges of the complete graph K_n there is a subset V_0 of vertices such that the difference between the number of red edges inside V_0 and the number of blue edges inside V_0 is at least $cn^{3/2}$, for some absolute constant $c > 0$. This problem is a convincing manifestation of the power of probabilistic methods, since they can be used in the other direction as well, to prove that the result is tight up to a constant factor. Additional examples of such results can be found in Alon and Spencer (2000).

4 Algorithmic Aspects and Future Challenges

As we have seen, it is one matter to prove that a certain combinatorial structure exists, and quite another to construct an example. A related question is whether an example can be generated by means of an EFFICIENT ALGORITHM [IV.20 §2.3], in which case we call it

explicit. This question has become increasingly important because of the rapid development of theoretical computer science, which has close connections with discrete mathematics. It is particularly interesting when the structures in question have been proved to exist by means of probabilistic arguments. Efficient algorithms for producing them are not just interesting on their own, but also have important applications in other areas. For example, explicit constructions of error-correcting codes that are as good as random ones are of major interest in CODING AND INFORMATION THEORY [VII.6], and explicit constructions of certain Ramsey-type colorings may have applications in DERANDOMIZATION [IV.20 §7.1.1] (the process of converting randomized algorithms into deterministic ones).

It turns out, however, that the problem of finding a good explicit construction is often very difficult. Even the simple proof of Erdős, described in section 3.2, that there are red/blue colorings of graphs with $\lfloor 2^{k/2} \rfloor$ vertices containing no monochromatic clique of size k leads to an open problem that seems very difficult. Can we construct, explicitly, such a graph with $n \geq (1 + \epsilon)^k$ vertices in time that is polynomial in n ? Here we allow ϵ to be any constant, as long as it is positive. This problem is still wide open, despite considerable efforts from many mathematicians.

The application of other advanced tools, such as algebraic and analytic techniques, spectral methods, and topological proofs, also tends to lead in many cases to nonconstructive proofs. The conversion of these to algorithmic arguments may well be one of the main future challenges of the area.

Another interesting recent development is the increased appearance of computer-aided proofs in combinatorics, starting with the proof of the FOUR-COLOR THEOREM [V.14]. To incorporate such proofs into the area, without threatening its special beauty and appeal, is a further challenge.

These challenges, the fundamental nature of the area, its tight connection to other disciplines, and its many fascinating open problems ensure that combinatorics will continue to play an essential role in the general development of mathematics and science in the future.

Further Reading

Alon, N., and J. H. Spencer. 2000. *The Probabilistic Method*, 2nd edn. New York: John Wiley.
 Bollobás, B. 1978. *Extremal Graph Theory*. New York: Academic Press.

Graham, R. L., M. Grötschel, and L. Lovász, eds. 1995. *Handbook of Combinatorics*. Amsterdam: North-Holland.
 Graham, R. L., B. L. Rothschild, and J. H. Spencer. 1990. *Ramsey Theory*, 2nd edn. New York: John Wiley.
 Janson, S., T. Łuczak, and A. Ruciński. 2000. *Random Graphs*. New York: John Wiley.
 Jukna, S. 2001. *Extremal Combinatorics*. New York: Springer.
 Matoušek, J. 2002. *Lectures on Discrete Geometry*. New York: Springer.
 Nathanson, M. B. 1996. *Additive Number Theory: Inverse Theorems and the Geometry of Sumsets*. New York: Springer.
 Pach, J., and P. Agarwal. 1995. *Combinatorial Geometry*. New York: John Wiley.
 Tao, T., and V. H. Vu. 2006. *Additive Combinatorics*. Cambridge: Cambridge University Press.

IV.20 Computational Complexity

Oded Goldreich and Avi Wigderson

1 Algorithms and Computation

This article is concerned with what can be computed efficiently, and what cannot. We will introduce several important concepts and research areas, such as formal models of computation, measures of efficiency, the \mathcal{P} versus \mathcal{NP} question, NP-completeness, circuit complexity, proof complexity, randomized computation, pseudorandomness, probabilistic proof systems, cryptography, and more. Underlying them all are the related notions of *algorithms* and *computation*, and we begin by discussing these.

1.1 What Is an Algorithm?

Suppose that you are presented with a large positive integer N and asked to determine whether it is prime. What should you do? One possibility would be to apply the method of *trial division*. That is, first see whether N is even, then whether it is a multiple of 3, then whether it is a multiple of 4, and so on through all the numbers up to \sqrt{N} . If N is composite, then it has a factor between 2 and \sqrt{N} , so it is prime if and only if the answer to all these questions is no.

The trouble with this method is that it is highly *inefficient*. Suppose, for instance, that N has 101 digits. Then \sqrt{N} is at least 10^{50} , so in order to carry the method out one would have to answer 10^{50} questions of the form, “Is K a factor of N ?” This would take far longer than a human lifetime, even if all the world’s computers devoted themselves to the task. What, then, is an “efficient procedure”? This question divides into two parts:

what is a procedure, and what counts as efficient? We shall look at these two questions in turn.

Three very obvious conditions that a method should satisfy if it is to count as a procedure for solving this problem are *finiteness*—that the procedure should have a finite description (so, for example, one cannot simply look up the answer in an infinite list of integers and their factorizations)—and *correctness*—that, for every N , it correctly tells you whether N is prime.

There is also a third, more subtle, condition, which goes to the heart of what is meant by the word “algorithm.” It is that it should consist of *simple steps*. This is needed in order to rule out ridiculous “procedures” such as, “See whether N has any nontrivial factors; declare N to be prime if and only if it does not.” The problem with this is that we cannot see, just like that, whether N has nontrivial factors. By contrast, all that the method of trial division asks of us is that we should do basic arithmetic, such as increasing integers by 1, comparing them, and doing long division. Moreover, the procedures of basic arithmetic can be broken down into yet simpler steps: for instance, it is possible to do long division by a succession of elementary operations applied to single digits at a time.

In order to understand this simplicity condition better, and to prepare ourselves for a formal definition of the notion of algorithms, let us look at long division in slightly more detail. Suppose that you have a piece of paper in front of you and you want to divide 5959578 by 857. You will write the two numbers down, and then, as the calculation proceeds, you will write other numbers as well. For instance, you may wish to start by writing out all the multiples of 857 up to 9×857 . At some point early on you will probably find yourself comparing $5999 = 7 \times 857$ with 5959: this you do by scanning the numbers from left to right and comparing individual digits. In this case, a difference is first detected in the third digit. You then write 5142 (which is 6×857) underneath the 5959, subtract (again by scanning numbers from left to right and performing single-digit operations), write down the difference 817, “bring down” the next digit, 5, of 5959578, and repeat the process with the number 8175.

At each stage in this calculation you are modifying the piece of paper in front of you. As you do so you need to keep track of which stage of the procedure you are at (whether you are writing out the initial table of multiples of 857, or seeing which one is the largest that does not exceed another number, or subtracting one number from another, or bringing down a digit, etc.),

and which symbols on the page you are currently dealing with. What is remarkable is that this information has a *fixed size*, in the sense that it does not increase as the size of the input (that is, the two numbers to be divided) increases.

Therefore, the procedure can be regarded as making *local changes* to some “environment,” using repeated applications of a fixed rule that does not depend on the input. (This rule will typically have some internal structure, such as a list of simpler rules together with specifications of the circumstances under which they should be applied.) In general, this is what we mean by a *computation*: it modifies an environment by means of repeated applications of a fixed rule. The rule is usually referred to as an *algorithm*. Notice that this description applies to many scientific theories of dynamic evolution in nature (of weather, chemical reactions, or biological processes, for example). Thus, these can be regarded as computational processes, of sorts. Some of these dynamical systems also demonstrate well the fact that simple, local rules can result in a very complex modification of the environment if they are iterated many times. (See DYNAMICS [IV.14] for further discussion of this phenomenon.)

Thoughts such as these lie behind the idea of a *Turing machine*, TURING’s [VI.94] famous formalization of the notion of an algorithm. It is interesting that he came up with his formalization *before* computers existed. Indeed, this abstraction and central features of it, most notably the existence of a “universal” machine, greatly influenced the actual construction of computers.

It is very important to know that the idea of an algorithm can be formalized, so that one can talk precisely about whether there are algorithms that will perform particular tasks, how many steps they need for a given size of input, and so on. However, there are many ways of doing this, which all turn out to be equivalent, and for the purposes of understanding this article it is not necessary to go into the details of any particular method. (You can, if you like, think of an algorithm as any procedure that can be programmed on a real computer—slightly idealized so that it has unlimited storage space—and a step of an algorithm as any change of one of the bits of that computer from a 0 to a 1 or vice versa.) Nevertheless, just to show roughly how it is done, here is a brief description of the basic features of the Turing machine model.

To begin with, one makes the observation that all computational problems can be encoded as operations on sequences of 0s and 1s. (This observation is not

just theoretically useful but also very important for the actual building of computers.) For example, all numbers that occur in the course of a computation can be converted into their binary representations; one can also use 1 to stand for “true” and 0 to stand for “false” and thereby perform the basic logical operations; and so on. For this reason we can define a very simple “environment” for a Turing machine: it is a “tape,” infinitely long in both directions, that consists of a row of “cells,” each of which contains either a 0 or a 1. Before the computation starts, a certain prespecified portion of this tape is filled with the *input*, which is a sequence of 0s and 1s. The algorithm is a little control mechanism. At any one time, this mechanism can be in one of a finite set of states, and it is located at one of the cells of the tape. According to the state it is in and the value, 0 or 1, that it sees at the cell it has reached, it makes three decisions: whether to change the value in the cell, whether to move left or right by one cell, and which state it should next be in.

One of the states of this control mechanism is “halt.” If this state is reached, then the mechanism stops doing anything and is said to have halted. At that point, a certain prespecified portion of the tape will be regarded as the output of the machine. An algorithm can be thought of as any Turing machine that halts for every possible input. And the number of steps of the algorithm is the number of steps taken by that Turing machine. Remarkably, this very simple computational model is enough to capture the full power of computation: in theory one could build a Turing machine, out of clockwork, say, that would be able to do whatever a modern supercomputer can do. (However, it would take too long over each step to be practical for anything but the very simplest of computations.)

1.2 What Does an Algorithm Compute?

A Turing machine converts a sequence of 0s and 1s into another sequence of 0s and 1s. If we wish to use mathematical language to discuss this, then we need to give a name to the set of $\{0, 1\}$ -sequences. To be precise, we consider the set of all *finite* sequences of 0s and 1s, and we call this set I . It is also useful to write I_n for the set of all $\{0, 1\}$ -sequences of length n . If x is a sequence in I , then we write $|x|$ for its length: for instance, if x is the string 0100101, then $|x| = 7$. To say that a Turing machine converts a sequence of 0s and 1s into another such sequence (if it halts) is to say that it naturally defines a function from I to I . If

M is the Turing machine and f_M is the corresponding function, then we say that M *computes* f_M .

Thus, every function $f : I \rightarrow I$ gives rise to a computational task, namely that of computing f . We say that f is *computable* if this is possible: that is, if there exists a Turing machine M such that the corresponding function f_M is equal to f . A central early result (due to Turing and independently to CHURCH [VI.89]) is that some natural functions are *not* computable. (For more details, see THE INSOLUBILITY OF THE HALTING PROBLEM [V.23].) However, complexity theory deals only with computable functions, and studies which of these can be computed *efficiently*.

Using the notation we have just introduced, we can formally describe various different kinds of computational tasks, of which two major examples are *search problems* and *decision problems*. The aim of a search problem is, informally speaking, to find a mathematical object with certain properties: for instance, one might wish to find a solution to a system of equations, and this solution might not be unique. We can model this by means of a binary RELATION [I.2 §2.3] R on the set I : for a pair (x, y) of strings in I , we say that y is a *valid solution of problem instance* x if xRy . (This notation means that x is related to y in the way specified by R ; another common notation for the same thing is $(x, y) \in R$.) For example, we might let x and y be binary expansions of positive integers N and K , respectively, and say that xRy if and only if N is a composite number and K is a nontrivial factor of N . Informally, this search problem would be, “Find a nontrivial factor of N .” If M is an algorithm that computes a certain function $f_M : I \rightarrow I$, then we say that M *solves the search problem* R if $f_M(x)$ is a valid solution of x for every problem instance x that has a solution. For example, it solves the search problem just defined if, for every composite number N with binary expansion x , $f_M(x)$ is the binary expansion of a nontrivial factor K of N .

Notice that in the above example we were interested in positive integers, but formally speaking an algorithm is a function of binary strings. This was not a problem, because there is a convenient and natural way to *encode* integers as binary strings—via their usual binary expansions. For the rest of this article, we shall feel free to blur the distinction between the mathematical objects we wish to investigate and the strings we use to represent them in a computation. For instance, it is simpler to think of the algorithm M in the previous paragraph as computing a function $f_M : \mathbb{N} \rightarrow \mathbb{N}$, and solving the search problem if, for every composite number N ,

$f_M(N)$ is a nontrivial factor of N . We stress that the representation of objects by strings is a rather succinct one: it takes only $\lceil \log_2 N \rceil$ bits to represent the number N , so the number N is exponentially larger than the length of its representation.

Now let us turn to decision problems. These are simply problems where one is looking for a yes/no answer. The problem with which we opened this article—Is N a prime number?—is a classic example of a decision problem. Notice that here and in the paragraph before last we are using the word “problem” in a slightly unusual way, to mean a general class of questions rather than just one. In this example, the question, “Is 443 a prime number?” would be called an *instance* of the problem, “Is N a prime number?”

Modeling decision problems is very simple: they are subsets of \mathbb{I} . The idea is that a subset S of \mathbb{I} consists of all the strings where the answer is yes. So if the problem is to determine primality, then S would consist of all binary expansions of prime numbers, at least if we chose the obvious encoding of the problem. When do we say that a machine M solves the decision problem S ? We would like it to compute a function f that says yes when the input x belongs to S and says no otherwise. That is, we say that M solves the problem S if the associated function f_M is a function from \mathbb{I} to the set $\{0, 1\}$ such that $f_M(x) = 1$ whenever $x \in S$ and $f_M(x) = 0$ otherwise.

Most of this article will be focused on decision problems, but the reader should bear in mind that computational tasks that seem more complicated, including search problems, can in fact usually be reduced to sequences of decision problems. For example, if you can solve all decision problems and you want to factorize a large composite number N , then you can proceed as follows. First, determine whether the smallest prime factor of the number ends in a 1 (in its binary expansion). If the answer is yes, you can look at the next digit by asking if this factor ends in 11; if it is no, then you can ask if it ends in 10. You can continue this process, extending your knowledge of the smallest prime factor by one bit at a time. The number of queries you will need to make will be at most the number of digits of N .

2 Efficiency and Complexity

Near the beginning of this article we asked what was meant by the phrase “efficient procedure.” We have now discussed the word “procedure” in some depth, but we have yet to say what we mean by “efficient,” beyond

pointing out that trial division takes too long to be practical if we have a very large integer and want to determine whether it is prime.

2.1 Complexity of Algorithms

How can we describe mathematically what it means for a procedure to “take too long to be practical”? The Turing-machine formalization is particularly useful for answering questions like this, because we can say precisely what a step of a Turing-machine computation is and this allows us to give a precise definition: an algorithm is a Turing machine, and its *complexity* is defined to be the number of steps the machine takes before halting.

If we look at this definition carefully, we see that what it defines is not just one number but a function. The time taken by a Turing machine depends on the input, so, given a Turing machine M and a string x , we can define $t_M(x)$ to be the number of steps M takes before halting when x is the input. The function $t_M : \mathbb{I} \rightarrow \mathbb{N}$ is the *complexity function* of the machine M .

Most of the time, we are interested not so much in the full detail of this complexity function, but in the *worst-case complexity* of the machine M . This is a function $T_M : \mathbb{N} \rightarrow \mathbb{N}$ defined as follows. Given a positive integer n , $T_M(n)$ is the maximum value of $t_M(x)$ over all input strings x of length n . In other words, we want to know the longest possible time that our machine might take when faced with an input of length n . And usually we do not look for an exact formula for $T_M(n)$: for most purposes it is enough to have a good upper bound.

The function $t_M(x)$ is more accurately called the *time complexity* of the algorithm M , since it measures how long M takes given x as its input. But time is not the only resource that matters in computer science. Another is how much memory an algorithm uses, beyond that needed to store the input, and this too can be captured in our formal model. Given a Turing machine M and an input x , we can define $s_M(x)$ to be the number of cells, other than input cells, that are visited before the machine halts, under the extra condition that the input cells must be left unchanged.

2.2 Intrinsic Complexity of Problems

Much of this article will be concerned with a very general analysis of the power of computation. In particular, we shall discuss a central subfield of theoretical computer science known as *computational complexity* (or *complexity theory*). The aim of this area is to

understand the *intrinsic complexity* of computational tasks.

Notice that we said “computational tasks” rather than “algorithms.” This is an important distinction and it involves a change of focus. Returning to our example of primality testing, it is not too hard to estimate how long various algorithms take, and indeed we had no trouble in seeing that trial division would take a very long time indeed. But does that mean that the task of primality testing is *intrinsically* hard? Not necessarily, since there may be other algorithms that do the job much more quickly.

This idea fits neatly into our formal scheme. What would be a good definition of the complexity of a computational task? Roughly speaking, the complexity of such a task should be the smallest complexity of any algorithm M that solves it. A convenient way of saying this is as follows. If $T : \mathbb{N} \rightarrow \mathbb{N}$ is some integer function, we say that the task has *complexity at most T* if there is an algorithm M that solves the task such that $T_M \leq T$ (i.e., $T_M(n) \leq T(n)$ for every n).

If you want to show that a computational task is not intrinsically hard, then all you have to do is devise an algorithm with low complexity that solves this task. But what if you want to show that this task is *intrinsically* hard? Then you have to prove, for *every possible* low-complexity algorithm M , that M does not solve this task. This is much harder: even after half a century of intensive work, the best results that are known are very weak. Notice a big difference between the two kinds of research: one can find algorithms without knowing how the concept of “algorithm” is formalized, but to analyze *all* algorithms with a certain property, it is essential to have a precise definition of what an algorithm is. Fortunately, with Turing’s formalization, we have one.

2.3 Efficient Computation and \mathcal{P}

Now we have ways of measuring the complexity of algorithms and computational tasks. But we have not yet addressed the question of when we should regard an algorithm as *efficient*, or a computational task as *efficiently solvable*. We shall propose a definition of efficiency that seems somewhat arbitrary and then explain why it is in fact a surprisingly good one.

If M is an algorithm, then we regard it as efficient if and only if it *terminates in polynomial time*. This means that there are constants c and k such that the worst-case complexity T_M always satisfies the inequality $T_M(n) \leq cn^k$. In other words, the time taken

by the algorithm is bounded above by a polynomial function of the length of the input string. It is not hard to convince yourself that the familiar methods for adding or multiplying two n -digit numbers terminate in polynomial time, whereas trial division for primality testing does not. Other familiar examples of tasks with efficient algorithms are putting a set of numbers in increasing order, computing the DETERMINANT [III.15] of a matrix (provided one uses row operations rather than substituting the entries directly into the formula), solving linear equations by Gaussian elimination, finding the shortest path in a given network, and more.

Since we are interested in the intrinsic complexity of computational tasks, we now define such a task to be *efficiently computable* if there is an efficient algorithm M that solves it. In our discussion of efficient computability, we shall focus on decision problems and consider the class of *all* decision problems that have efficient algorithms. Understanding it is *the* major goal of computational complexity theory. Here is a formal definition. We shall use the following convenient piece of notation: if M is a Turing machine and x is an input, then $M(x)$ is the output of x . (Earlier we wrote $f_M(x)$ for this function.) Since we are considering decision problems, $M(x)$ will be 0 or 1.

Definition. A decision problem $S \subseteq I$ is *solvable in polynomial time* if there is a Turing machine M , terminating in polynomial time, such that $M(x) = 1$ if and only if $x \in S$.

The class of decision problems that are solvable in polynomial time is our first example of a *complexity class*. It is denoted \mathcal{P} .

The *asymptotic analysis* of running time, i.e., estimating the running time as a function of the input length, turns out to be crucial for revealing structure in the theory of efficient computation. The choice of polynomial time as the standard for efficiency may seem arbitrary, and theories could be developed with other choices, but it has amply justified itself. The main reason for this is that the class of polynomials (or functions bounded above by a polynomial) is closed under various operations that arise naturally in computation. In particular, the sum, product, or composition of two polynomials is again a polynomial. This allows us, for example, to think of long division as a basic, one-step operation when we are investigating the efficiency of algorithms for primality testing. In fact, long division takes more than one step, but it is in \mathcal{P} so the time it

takes does not affect whether an algorithm that uses it is itself in \mathcal{P} . In general, if we use the basic programming technique of *subroutines*, and if our subroutines are in \mathcal{P} , then we will preserve the efficiency of the algorithm as a whole.

Almost all computer programs that are used in practice turn out to be efficient in this theoretical sense. Of course, the converse is not true: an algorithm that runs in time n^{100} is completely useless despite the fact that n^{100} is a polynomial. However, this seems not to matter. It is unusual to discover even an n^{10} -time algorithm for a natural problem, and on the rare occasions when this happens, improvements to n^3 - or n^2 -time, which border on the practical, almost always follow.

It is important to contrast \mathcal{P} with the class \mathcal{EXP} . A problem belongs to \mathcal{EXP} if there is an algorithm that solves it in at most $\exp(p(n))$ steps for any input of length n , where p is some polynomial. (Roughly speaking, \mathcal{EXP} consists of problems that can be solved in exponential time: the polynomial p makes the definition more robust and less dependent on the precise nature of encodings, etc.)

If you use trial division to test the primality of a number N with n digits in its binary expansion, then you have to do \sqrt{N} long-division calculations. Since \sqrt{N} is about $2^{n/2}$, this is an exponential-time procedure. Exponential running time is considered blatantly *inefficient*, and if the problem has no faster algorithm, then it is deemed intractable. It is known (via a basic technique called *diagonalization*) that $\mathcal{P} \neq \mathcal{EXP}$; furthermore, some problems in \mathcal{EXP} really do require exponential time. Almost all problems and classes considered in this paper can easily be shown to belong to \mathcal{EXP} via trivial, “brute-force” algorithms such as the trial division just discussed: the main question will be whether much faster algorithms can be devised for them.

3 The \mathcal{P} versus \mathcal{NP} Question

In this section we discuss the famous \mathcal{P} versus \mathcal{NP} question, which is usually formulated in terms of decision problems, but which also has an interpretation in terms of search problems. We shall start with the latter.

3.1 Finding versus Checking

Can you rearrange the letters CHAIRMITTE to form an English word? To solve a puzzle like this, one has to search among many possibilities (all permutations of those letters), perhaps building up fragments of words and hoping that inspiration will strike. Now consider

the following question: can the letters of CHAIRMITTE be rearranged to form the word “arithmetic”? It is very easy (if slightly boring) to check that the answer is yes.

This informal example illustrates an important feature of many search problems: that once you find a solution, it is easy to recognize that it *is* a solution. The hard part is to find the solution in the first place. Or at least, so it seems. But actually proving that search problems of this kind are hard is a famous unsolved problem, the \mathcal{P} versus \mathcal{NP} question.

Another search problem with this quality, which is in fact quite general and has a natural appeal to mathematicians, is the task of finding proofs for valid mathematical statements. Again it seems to be far easier to check that an argument is a valid proof than it is to find the argument in the first place. Since finding a proof is a process that requires considerable creativity (as, in a much smaller way, is finding an anagram), the \mathcal{P} versus \mathcal{NP} question is, in a sense, asking whether this kind of creativity can be automated.

In section 3.2 we shall define the class \mathcal{NP} formally. Informally, it corresponds to the set of all search problems for which it is easy to check whether you have found what you are searching for. Another example of such a problem is that of finding a factor of a large composite integer N . If you are told that K is a factor, then it is an easy task for you (or your computer) to verify that this is true: all you have to do is a single instance of long division.

A vast number of problems in science (such as creating theories to explain various natural phenomena) and engineering (such as creating designs under various physical and economic constraints) have the same property that success is much easier to recognize than to achieve in the first place. This gives some indication of the importance of this class of problems.

3.2 Deciding versus Verifying

For the purposes of theoretical analysis, it is actually more convenient to define \mathcal{NP} as a class of *decision* problems. For instance, consider the decision problem, “Is N composite?” What makes this a problem in \mathcal{NP} is that, whenever N is composite, there is a *short proof* of this fact. Such a proof consists of a factor of N , and is easy to check that this proof is correct. That is, it is easy to devise a polynomial-time algorithm M that takes as input a pair (N, K) of positive integers and outputs 1 if K is a nontrivial factor of N and 0 otherwise. If N is prime, then $M(N, K) = 0$ for every K , while

if N is composite there will always exist an integer K such that $M(N, K) = 1$. Moreover, in this case the string that encodes K will be at most as long as the string that encodes N , though all we really care about is that it should not be too much longer. These properties we now encapsulate in a formal definition.

Definition (the complexity class \mathcal{NP}^1). A decision problem $S \subset \mathbf{I}$ belongs to \mathcal{NP} if there is a subset $R \subset \mathbf{I} \times \mathbf{I}$ with the following three properties.

- (i) There is a polynomial function p such that $|y| \leq p(|x|)$ whenever $(x, y) \in R$.
- (ii) x belongs to S if and only if there is some y such that (x, y) belongs to R .
- (iii) The problem of determining whether a pair (x, y) belongs to R is in \mathcal{P} .

When such a y exists, it is called a *proof* (or *witness*) of the fact that x belongs to S . The polynomial-time algorithm for determining whether a pair (x, y) belongs to R is called a *verification procedure* for determining whether x belongs to S .

Notice that every problem S in the class \mathcal{P} is also in \mathcal{NP} , since we can simply forget about the candidate proof y and use the efficient test for whether x belongs to S . On the other hand, every problem in \mathcal{NP} is trivially in \mathcal{EXP} , because we can enumerate all possible y s (in exponential time) and check for each one whether it works. (This is more or less what we do with trial division.) Can this trivial algorithm be improved? Sometimes it can, even in very nonobvious cases. In fact, recently it was proved that the problem of determining whether a number N is composite belongs to \mathcal{P} . (Further details can be found in COMPUTATIONAL NUMBER THEORY [IV.3 §2].) However, we would like to know whether for *every* problem in \mathcal{NP} one can do much better than the trivial algorithm.

3.3 The Big Conjecture

The \mathcal{P} versus \mathcal{NP} problem asks whether or not \mathcal{P} equals \mathcal{NP} . In terms of decision problems, this question is asking *whether the existence of an efficient verification procedure for some set implies the existence of an efficient decision procedure for it*. In other words, if there is a polynomial-time algorithm for checking

whether proofs that $x \in S$ are correct (as in the definition of \mathcal{NP} just given), does it follow that there is a polynomial-time algorithm for deciding whether $x \in S$?

As our earlier examples suggest, the problem can also be formulated as a question about search problems. Suppose we have a set $R \subset \mathbf{I} \times \mathbf{I}$ satisfying properties (i) and (iii) of the definition of \mathcal{NP} . For instance, R might correspond to all pairs of integers (N, K) such that K is a nontrivial factor of N . Then the corresponding search problem, “Given a composite number N find a nontrivial factor K ,” is closely related to the integer factorization problem. In general, any such relation R gives rise to a search problem, “Given a string x , find a string y such that (x, y) belongs to R (if such a y exists).” Now the \mathcal{P} versus \mathcal{NP} problem asks the following: “Are all such search problems solvable in polynomial time?”

If the answer is yes, then the *mere fact* that it can be checked in polynomial time whether K is a nontrivial factor of N would imply that such a factor could actually be found in polynomial time.² Similarly, the mere fact that a short proof of a mathematical statement existed would be enough to guarantee that it could be found in a short time by a purely mechanical process. The apparent difference between the difficulty of discovering solutions and the ease of checking them once discovered would be entirely illusory.

This would be very strange, and almost all experts believe that it is not the case. However, nobody has managed to prove it. So the big conjecture is that \mathcal{P} does not equal \mathcal{NP} . That is, finding is harder than checking, and efficient verification procedures do not necessarily lead to efficient algorithms for decision problems. This conjecture is strongly supported by our intuition, which has been developed over many centuries of dealing with search and decision problems in a wide variety of human activities. Further empirical evidence in favor of the conjecture is given by the fact that there are literally thousands of \mathcal{NP} problems, from many mathematical and scientific disciplines, that are not known to be solvable in polynomial time, despite the fact that researchers have tried very hard to discover efficient procedures for solving them.

The $\mathcal{P} \neq \mathcal{NP}$ conjecture is certainly the most important open problem in computer science, and one of the most significant in all of mathematics. Our later section on circuit complexity (section 5.1) is devoted

1. The acronym \mathcal{NP} stands for nondeterministic polynomial-time, where a *nondeterministic machine* is a *fictitious* computing device used in an alternative definition of the class \mathcal{NP} . The nondeterministic moves of such a machine correspond to guessing a “proof” in this definition.

2. Despite the fact that there is a polynomial-time algorithm for determining whether a number is composite, no such algorithm is known for actually finding its factors, and it is widely believed that no efficient algorithm exists for this.

to attempts to prove it. There we shall discuss some partial results and limits of the techniques used so far.

3.4 \mathcal{NP} versus $\text{co } \mathcal{NP}$

Another important class, known as $\text{co } \mathcal{NP}$, is the class of *complements* of sets (or decision problems) in \mathcal{NP} . For example, the problem “Is N prime?” belongs to $\text{co } \mathcal{NP}$ because there is an efficient verification procedure for showing that a given positive integer N is *not* prime, namely, exhibiting some factors. Equivalently, the set of primes belongs to $\text{co } \mathcal{NP}$ because its complement belongs to \mathcal{NP} .

Does \mathcal{NP} equal $\text{co } \mathcal{NP}$? That is, if you have an efficient verification procedure for determining membership of a set S , do you also have one for determining *nonmembership*? Again, intuition would suggest not, or at least not necessarily. For instance, if a jumble of letters can be rearranged to form a word, then that word serves as a short demonstration. But suppose a jumble of letters *cannot* be rearranged to form a word. One could demonstrate this by looking at all possible rearrangements and noting that none of them is a word, but this is a very long demonstration and there does not seem to be a systematic way of finding a truly short one.

Here again intuition from mathematics is extremely relevant: to verify that a set of logical constraints is mutually *inconsistent*, that a family of polynomial equations has *no* common root, or that a set of regions in space has *empty* intersection seems far harder than to verify the opposite (exhibiting a consistent valuation, a common root, or a point that belongs to all the regions). Indeed, only when rare extra mathematical structure is available, such as DUALITY [III.19] theorems or complete systems of invariants, are we able to show that a set and its complement are computationally equivalent. So another big conjecture is that \mathcal{NP} is not equal to $\text{co } \mathcal{NP}$. The section on proof complexity (section 5.3) looks further at this conjecture and at attempts to resolve it.

Surprisingly, it is not hard to show that the problem, “Is N composite?” which obviously belongs to \mathcal{NP} , actually belongs to $\text{co } \mathcal{NP}$ as well. To prove this, one uses the following fact from elementary number theory: p is prime if and only if there is an integer $a < p$ such that $a^{p-1} \equiv 1 \pmod{p}$ and $a^r \not\equiv 1$ whenever r is a factor of $p - 1$. Thus, to verify that p is prime it is enough to exhibit such an integer a . However, to check that a works, one needs to know the prime factorization of $p - 1$, and one must give a short proof that it

really is a factorization into primes. This takes us back to the problem we started with, but the numbers are smaller so one can give a recursive argument. (We mention again that the set of primes is actually in \mathcal{P} , but this is harder to prove.)

4 Reducibility and NP-Completeness

One sign that a mathematical problem is fundamental is that it has many equivalent formulations. This is true to a quite extraordinary extent for the \mathcal{P} versus \mathcal{NP} problem, as we shall see in this section. Fundamental to our discussion will be the notion of *polynomial-time reducibility*. Roughly speaking, one computational problem is polynomially reducible to another if any polynomial-time algorithm for the second can be converted into a polynomial-time algorithm for the first. Let us see an example of this, and then we will define the notion formally.

First, here is a famous problem in \mathcal{NP} , called SAT. Consider the logical formula

$$(p \vee q \vee \bar{r}) \wedge (\bar{p} \vee q) \wedge (p \vee \bar{q} \vee r) \wedge (\bar{p} \vee \bar{r}).$$

Here, p , q , and r are *propositions*, each of which can be true or false. The symbols “ \vee ” and “ \wedge ” stand for OR and AND, respectively, and \bar{p} (read as “NOT- p ”) is the proposition that is true if and only if p is false.

Suppose now that p is true, q is true, and r is false. Then the first subformula $p \vee q \vee \bar{r}$ is true because at least one of p , q , and \bar{r} is true. Similarly, one can check that all the other subformulas are true, which means that the entire formula is true. We call our choice of truth values for p , q , and r a *satisfying assignment* for the formula, and we say that the formula is *satisfiable*. A natural computation problem that arises is the following.

SAT: *given a propositional formula, is it satisfiable?*

In the example above, the formula was a conjunction of subformulas, called *clauses*. In their turn, these subformulas were disjunctions of propositions or their negations, which are called *literals*. (The *conjunction* of some formulas ϕ_1, \dots, ϕ_k is the formula $\phi_1 \wedge \dots \wedge \phi_k$ and their *disjunction* is $\phi_1 \vee \dots \vee \phi_k$.)

3SAT: *given a propositional formula that consists of a conjunction of clauses that contain at most three literals each, is the formula satisfiable?*

Notice that SAT and 3SAT are in \mathcal{NP} , since it is an easy matter to check whether a given truth assignment to the variables is a satisfying assignment for the formula.

Let us now turn to a second problem in \mathcal{NP} .

3-colorability: *given a planar map (such as one might find in an atlas), can its regions be colored with three colors, Red, Blue, and Green, such that no two adjacent countries have the same color?*³

We shall now “reduce” 3-colorability to 3SAT: that is, show how an algorithm that solves 3SAT can be used to solve 3-colorability as well. Suppose, then, that we have a map with n regions. We shall need $3n$ propositions, which we shall call R_1, \dots, R_n , B_1, \dots, B_n , and G_1, \dots, G_n , and we would like to define a logical formula in such a way that a satisfying assignment of the formula will correspond to a 3-coloring of the graph. In the back of our minds, we shall think of R_i as the statement, “Region i of the map is colored Red,” and similarly for B_i and G_i . We then take as our clauses some statements that tell us that every region receives a single color and no two adjacent regions receive the same color.

This is easy to do: to guarantee that region i receives a color, we take the clause $R_i \vee B_i \vee G_i$, and if regions i and j are adjacent, then to guarantee that they do not receive the same color we take the three clauses $\overline{R_i} \vee \overline{R_j}$, $\overline{B_i} \vee \overline{B_j}$, and $\overline{G_i} \vee \overline{G_j}$. (To ensure that no region is assigned more than one color, we can also add clauses of the form $\overline{R_i} \vee \overline{B_i}$, $\overline{B_i} \vee \overline{G_i}$, and $\overline{G_i} \vee \overline{R_i}$. Alternatively, we can allow multiple colors and finish by picking one of the assigned colors for each region.)

It is not hard to see that the conjunction of all these clauses is satisfiable if and only if there is a 3-coloring of the map. Furthermore, the conversion process is a simple one that can be carried out in a time that is polynomial in the number of regions in the map. Thus, we have our hoped-for polynomial-time reduction.

Now let us give a formal description of what we have just done.

Definition (polynomial-time reducibility). Let S and T be subsets of \mathbf{I} . We say that S is *polynomial-time reducible* to T if there exists a polynomial-time computable function $h : \mathbf{I} \rightarrow \mathbf{I}$ such that $x \in S$ if and only if $h(x) \in T$.

If S is polynomial-time reducible to T , then the following algorithm can be used to decide membership of S : given x , compute $h(x)$ (in polynomial time), then decide whether $h(x) \in T$. Therefore, if membership of T can be decided in polynomial time, so can membership of S . An equivalent, and important, way of saying

this is that if membership of S cannot be decided in polynomial time, then neither can membership of T . In short, if S is hard, then T is hard.

Now let us give a very important definition based on the notion of polynomial-time reducibility.

Definition (NP-completeness). A decision problem S is *NP-complete* if S is in \mathcal{NP} and every decision problem in \mathcal{NP} is polynomial-time reducible to S .

That is, if S has a polynomial-time algorithm, then so do *all other* problems in \mathcal{NP} . Thus, an NP-complete (decision) problem is in a certain sense “universal” among all problems in \mathcal{NP} .

At first this may seem a peculiar definition, because it is far from obvious that there are any NP-complete problems! However, in 1971, it was proved that SAT is NP-complete, and since then thousands of problems have been proved to be NP-complete as well. (Hundreds of them are listed in Garey and Johnson (1979).) Other examples are 3SAT and 3-colorability. The significance of 3SAT is that it is one of the most basic of all NP-complete problems. (It is not too hard to show that, by contrast, 2SAT and 2-colorability have polynomial-time algorithms.) In order to prove that a decision problem S is NP-complete, one starts with a known NP-complete problem S' and finds a polynomial-time reduction from S' to S . It now follows that if S has a polynomial-time algorithm, then so does S' and hence so do all other problems in \mathcal{NP} . Sometimes these reductions are quite simple, like our reduction of 3-colorability to 3SAT. But sometimes they need a great deal of ingenuity.

Here are two further NP-complete problems.

Subset sum: *given a sequence of integers a_1, \dots, a_n and another integer b , does there exist a set J such that $\sum_{i \in J} a_i = b$?*

Traveling salesman problem: *given a finite GRAPH [III.34] G , does there exist a Hamilton cycle? That is, can one find a cycle of edges that visits each vertex of the graph exactly once?*

Interestingly, almost all natural problems in \mathcal{NP} that are not obviously in \mathcal{P} turn out to be NP-complete. However, there are two important examples that have not been shown to be NP-complete and are strongly believed not to be. The first is a problem we have already discussed: integer factorization. More precisely, consider the following decision problem.

3. Recall that the celebrated FOUR-COLOR THEOREM [V.14] asserts that this can always be done with four colors.

PUP: Tim thinks that if you've followed the article to this point, then it's clear what this sentence means. OK to leave it as it is?

Factor in interval: *given x, a, b , does x have a prime factor y such that $a \leq y \leq b$?*

A polynomial-time algorithm for this can be combined with a simple binary search to find a prime factor if it exists. The reason this problem is unlikely to be NP-complete is that it also belongs to coNP . (Roughly speaking, this is true because one can exhibit the prime factorization of x and demonstrate in polynomial time that it really is a prime factorization.) If it were NP-complete, then it would follow that $\text{NP} \subset \text{coNP}$, and hence, by symmetry, that $\text{NP} = \text{coNP}$.

The second example is the following.

Graph isomorphism: *given two graphs G and H with n vertices, is there a function ϕ from the vertex set of G to the vertex set of H such that $\phi(x)\phi(y)$ is an edge of H when, and only when, xy is an edge of G ?*

Notice with these two examples how surprising it is that they can be reduced in polynomial time to problems such as 3SAT or 3-colorability. This is particularly true of the first, which has nothing to do with graphs or satisfiability of logical formulas.

If $\text{P} \neq \text{NP}$, then no NP-complete problem has a polynomial-time decision procedure. Consequently, the corresponding search problems cannot be solved in polynomial time. Thus, a proof that a problem is NP-complete is often taken as *evidence* that this problem is hard: if we could solve it, then we could also efficiently solve a multitude of other problems. But thousands of researchers (and tens of thousands of engineers) have, over several decades, tried and failed to find such procedures.

NP-completeness has more positive aspects as well. Sometimes it is possible to prove a fact about all sets in NP by establishing it only for some NP-complete set (and noting that polynomial-time reductions preserve the claimed property). Famous examples include the existence of “zero-knowledge proofs,” established first for 3-coloring (see section 6.3.2), and the so-called *PCP theorem*, established first for 3SAT (see section 6.3.3).

5 Lower Bounds

As we mentioned earlier, it is very much harder to prove that certain problems *cannot* be solved efficiently than it is to find efficient algorithms (when they exist). In this section, we shall survey some of the basic methods that have been developed for finding lower bounds for the complexity of natural computational problems. That is,

we shall discuss results that say that no algorithm can run in fewer than a given number of steps.

In particular, we shall introduce the theories of *circuit complexity* and *proof complexity*. The first is defined with the long-term goal of proving that $\text{P} \neq \text{NP}$, and the second is a program that is aimed at proving that $\text{NP} \neq \text{coNP}$. Both of these theories use the notion of a *directed acyclic graph*, which models the flow of information in a computation or a proof, and the sequence of derivations of each new piece of information from previous ones.

A directed graph is a graph for which each edge is given a direction. One can visualize it as a graph with arrows along the edges. A *directed cycle* is a sequence of vertices v_1, \dots, v_t such that for every i between 1 and $t-1$ there is an edge pointing from v_i toward v_{i+1} and there is also an edge pointing from v_t back to v_1 . If a directed graph G has no directed cycle, then it is called *acyclic*. We shall abbreviate the phrase “directed acyclic graph” by writing DAG.

It is not hard to see that in every DAG there will be some vertices with no incoming edges and some with no outgoing edges. These are called *inputs* and *outputs*, respectively. If u and v are vertices of a DAG and there is an edge from u to v , then we say that u is a *predecessor* of v . The basic idea of the DAG model is that you place information at each input, and at each vertex v you have a very simple rule that derives some information at v from the information at all the predecessors of v . Starting at the inputs, you gradually move through the graph, working out the information at a vertex once you have worked out the information for all its predecessors, until you have reached all the outputs.

5.1 Boolean Circuit Complexity

A Boolean circuit is a DAG in which all the values at the inputs, outputs, and intermediate vertices are *bits*. That is, each vertex may take the value 0 or 1. We have to specify simple rules for determining the value at a vertex from the values of its predecessors, and the usual choice is to allow three logical operations: AND, OR, and NOT. We call a vertex v an *AND gate* if the following rule applies: the value at v is 1 if all its predecessors have value 1 and is otherwise 0. At an *OR gate* we have a similar rule: the value at v is 1 if and only if at least one of its predecessors has value 1. Finally, v is a *NOT gate* if it has exactly one predecessor u , and v takes the value 1 if and only if u takes the value 0.

Given any Boolean circuit with n inputs u_1, \dots, u_n and m outputs v_1, \dots, v_m one can associate with it a function f from \mathbf{I}_n to \mathbf{I}_m as follows. Given a $\{0, 1\}$ -string $x = (x_1, \dots, x_n)$ of length n , let each u_i take the value x_i . Then use the gates of the circuit to find the values at the outputs v_1, \dots, v_m . If these are y_1, \dots, y_m , then $f(x_1, \dots, x_n) = (y_1, \dots, y_m)$.

It is not hard to prove that *any* function from \mathbf{I}_n to \mathbf{I}_m can be computed in this way. Thus, we say that AND, OR, and NOT gates, or more briefly “ \wedge ”, “ \vee ”, and “ \neg ”, form a *complete basis*. Moreover, this is true even if we restrict attention to DAGs where every vertex has at most two predecessors. In fact, we shall now assume that our DAGs have this property unless we say otherwise. There are other choices of gates that are complete bases, but we shall stick with “ \wedge ”, “ \vee ”, and “ \neg ” since this does not affect our discussion in an essential way.

It may be easy to show that every Boolean function f can be computed by means of a circuit, but as soon as one asks how large the circuit needs to be, one comes up against fascinating and very difficult questions. Thus, the following definition is central to the subject of circuit complexity.

Definition. Let f be a function from \mathbf{I}_n to \mathbf{I}_m . Then $S(f)$ is the size of the smallest Boolean circuit that computes f , where this is measured by the number of vertices in the corresponding DAG.

To see what this has to do with the \mathcal{P} versus \mathcal{NP} question, consider an NP-complete decision problem such as 3SAT. This can be coded as a function f from \mathbf{I} to $\{0, 1\}$, with $f(x)$ taking the value 1 if and only if the formula corresponding to x is satisfiable. Now we cannot find a circuit to compute f for the simple reason that \mathbf{I} is an infinite set. However, if we restrict attention to formulas that can be encoded as strings of length n , then we obtain a function $f_n : \mathbf{I}_n \rightarrow \{0, 1\}$, and we can try to estimate $S(f_n)$.

If we do this for every n , then we obtain an estimate for the growth rate of $S(f_n)$ as n tends to infinity. Writing f for the infinite sequence of functions (f_1, f_2, \dots) , let us define $S(f)$ to be the function that takes n to $S(f_n)$.

This is an important definition because of the following fact: if there is a polynomial-time algorithm for computing f , then the function $S(f)$ is bounded above by a polynomial. More generally, given any function $f : \mathbf{I} \rightarrow \mathbf{I}$, let f_n stand for the restriction of f to \mathbf{I}_n . If f has Turing complexity T (as defined in section 2.1), then $S(f_n)$ is bounded above by a polynomial function of

$T(n)$. That is, there is a sequence of circuits that computes the function f , and takes a time not significantly different from the time taken by the Turing machine.

This provides us with a potential method of proving lower bounds on computational complexity, since if we can prove that $S(f_n)$ grows very rapidly with n , then we have proved that the Turing complexity of f is very large. If f is a problem in \mathcal{NP} , then this proves that $\mathcal{P} \neq \mathcal{NP}$.

The circuit model of computation is finite rather than infinite, which raises an issue called *uniformity*. When we build a family of circuits from a Turing machine, the circuits are all in a certain sense “the same.” More precisely, there is an algorithm that can generate these circuits, and the time it takes to generate each one is polynomial in its size. A uniform family of circuits is one that can be generated in this way.

However, by no means all families of circuits are uniform. Indeed, there are functions f that cannot be computed by Turing machines at all (let alone in a reasonable amount of time), despite having circuits of *linear* size. This extra power comes from the fact that these families of circuits do not have a succinct (“effective”) description; that is, there is no single algorithm that can generate them. Such families are called *nonuniform*.

If there are many families of circuits that do not arise from Turing machines, then it would seem that proving good lower bounds for circuit complexity should be much harder than proving lower bounds for Turing complexity, since now one must rule out many more potential ways of computing a function. However, there is a strong sentiment that the extra power provided by nonuniformity is irrelevant to the \mathcal{P} versus \mathcal{NP} question: it is believed that for a natural problem such as 3SAT, nonuniformity does not help. Therefore, we have another big conjecture of theoretical computer science: that NP-complete sets do not have polynomial-size circuits. Why do we believe this conjecture? It would be nice to be able to say that its falsehood implied that $\mathcal{P} = \mathcal{NP}$.

We do not quite know that, but we do know that if it is false then “the polynomial-time hierarchy collapses.” Roughly speaking, this means that a whole system of complexity classes, which appear to be distinct, would in fact all be the same, which would be very unexpected. In any case, it is hard to imagine that there might be a sequence of polynomial-sized circuits computing an NP-complete problem without its being possible to generate such a sequence by an efficient algorithm.

Even if we grant that nonuniformity does not help solve NP-complete problems, what is the point of replacing the Turing machine model by the more powerful model of circuit families? The main reason is that circuits are simpler mathematical objects than Turing machines, and have the great advantage of being *finite*. The hope is that, while abstracting away the uniformity condition, which ought to be irrelevant, circuits provide us with a model that can be analyzed using combinatorial techniques.

It is also worth mentioning that Boolean circuits are a natural computational model of “hardware complexity,” so their study is of independent interest. Moreover, some of the techniques for analyzing Boolean functions have found applications elsewhere: for example, in computational learning theory, combinatorics, and game theory.

5.1.1 Basic Results and Questions

We have already mentioned several basic facts about Boolean circuits, in particular the fact that they can efficiently simulate Turing machines. Another basic fact is that *most Boolean functions require exponential-size circuits*. This can be proved by a simple counting argument: the number of small circuits is far smaller than the number of functions. More precisely, let the number of inputs be n . The number of possible functions defined on the set of all n -bit sequences is precisely 2^{2^n} . On the other hand, it is not hard to show that the number of circuits of size m is bounded above by around m^{m^2} . It follows easily that we cannot compute all functions unless $m > 2^{n/2}/n$. Furthermore, the proportion of functions that can be computed by a circuit of size at most m is tiny.

Thus, hard functions (for circuits and consequently for Turing machines) abound. However, this hardness is proved via a counting argument, which does not give us a way of actually exhibiting a hard function. That is, we cannot prove such hardness for any *explicit* function f , where “explicit” means that we place some algorithmic restriction on f , such as belonging to \mathcal{NP} or \mathcal{EXP} . In fact, the situation is even worse: no *nontrivial* lower bound is known for any explicit function. For any function f on n bits (assuming that it depends on all its inputs), we trivially must have $S(f) \geq n$, just to read the inputs. A major open problem of circuit complexity is beating this trivial bound by more than a constant factor.

Open problem. Find an explicit Boolean function f (or even a length-preserving function f) for which $S(f)$ is superlinear: that is, not bounded above by cn for any constant c .

A particularly basic special case of this problem is the question of whether addition is easier than multiplication. Let ADD and MULT denote, respectively, the addition and multiplication functions defined on pairs of integers (presented in binary). For addition, the usual procedure one learns at school gives rise to a linear-time algorithm, which implies a linear upper bound for $S(\text{ADD})$ as well. For multiplication, the standard school algorithm runs in *quadratic* time: that is, the number of steps is proportional to n^2 . This can be greatly improved (via FAST FOURIER TRANSFORMS [III.26]) to an algorithm that yields $S(\text{MULT}) < n(\log n)^2$. Since $\log n$ grows very slowly with n , this is only slightly superlinear. And now the question is whether this can be improved further. In particular, do there exist linear-size circuits for multiplication?

How can circuit complexity be a thriving subject if no nontrivial bounds are known for any explicit functions? The answer is that there have been some remarkable successes in proving lower bounds under natural extra assumptions on the circuits. We shall now describe the most important of these extra assumptions.

5.1.2 Monotone Circuits

As we have seen, general Boolean circuits can compute every Boolean function, and can do it at least as efficiently as general algorithms. Now some functions have additional properties that might lead one to expect that they could be computed with Boolean circuits of a particular kind. For example, consider the function CLIQUE, defined on the set of all graphs as follows. If G is a graph with n vertices, then a *clique* in G is defined to be a set of vertices such that any two are joined by an edge. Let us define $\text{CLIQUE}(G)$ to be 1 if G contains a clique of size at least \sqrt{n} and 0 otherwise.

Notice that if we add an edge to G , then either $\text{CLIQUE}(G)$ changes from 0 to 1 or it stays the same. What it will *not* do is change from 1 to 0: adding an edge obviously cannot *destroy* a clique.

We can encode G as a string x of $\binom{n}{2}$ bits, one for each pair of vertices, assigning 1 to a bit if the corresponding pair of vertices is joined by an edge and 0 otherwise. If we then set $\text{CLIQUE}(x)$ to equal $\text{CLIQUE}(G)$, we find that changing any bit of x from a 0 to a 1 cannot

change $\text{CLIQUE}(x)$ from 1 to 0. Boolean functions with this property are called *monotone*.

When considering the complexity of monotone functions, it is extremely natural to restrict the circuits by allowing only AND and OR gates, and disallowing NOT gates. Notice that “ \wedge ” and “ \vee ” are monotone operations, in the sense that changing an input bit from 0 to 1 will not change the output of the gate from 1 to 0, whereas “ \neg ” is certainly not monotone in this sense. A circuit that uses just “ \wedge ” and “ \vee ” is called a *monotone circuit*. It is not hard to show that every monotone function $f : \mathbb{I}_n \rightarrow \mathbb{I}_m$ can be computed by a monotone circuit, and that almost all monotone functions need exponential-sized circuits.

Does the extra restriction on the circuits make it easier to prove lower bounds? For over forty years the answer seemed to be not much: nobody could prove a super-polynomial lower bound for the monotone complexity of any explicit monotone function. But then, in 1985, a new technique called the *approximation method* was invented to prove the remarkable theorem that CLIQUE has super-polynomial monotone complexity. This technique eventually led to the following even stronger result.

Theorem. *CLIQUE requires monotone circuits of exponential size.*

Very roughly speaking, the approximation method works as follows. Assume that CLIQUE can be computed with a small monotone circuit. Then replace the occurrences of “ \wedge ” and “ \vee ” in this circuit with other gates that are cleverly chosen (and complex to describe), denoting these by “ $\hat{\wedge}$ ” and “ $\hat{\vee}$,” respectively. The new gates are chosen to satisfy two key properties.

- (i) Replacing one particular gate has only a “small” effect on the output of the circuit (where “small” is defined in terms of a certain natural but nontrivial measure of distance). Consequently, if a circuit has few gates, then replacing all of them yields a new circuit that approximates the original circuit for “most” choices of inputs.
- (ii) On the other hand, *every* circuit (regardless of its size) containing only the approximating gates “ $\hat{\wedge}$ ” and “ $\hat{\vee}$ ” computes a function that can be shown to be “far” from CLIQUE , in the sense that it disagrees with CLIQUE on many inputs.

CLIQUE is a well-known NP-complete problem, so the above theorem provides us with an explicit monotone

function, conjectured not to be in \mathcal{P} , that cannot be computed by small monotone circuits. It is natural at this point to wonder whether every monotone function that *is* in \mathcal{P} can be computed by a small monotone circuit. If so, we would be able to deduce that $\mathcal{P} \neq \mathcal{NP}$. However, the same method yields a *super-polynomial* lower bound for the size of monotone circuits that compute the PERFECT MATCHING function, which is monotone and is in \mathcal{P} . Given a graph G , this function outputs 1 if one can pair up the vertices in such a way that every pair is connected by an edge and 0 otherwise. Furthermore, exponential-size lower bounds are known for other monotone functions in \mathcal{P} , so general circuits are known to be substantially more powerful than monotone circuits, even for computing monotone functions.

5.1.3 Bounded-Depth Circuits

To understand the motivation for our next model, consider the following basic question: “Can one speed up computation by using several computers in parallel?” For instance, suppose that a certain task can be performed by one computer in t steps. Can it be performed by t (or even t^2) cooperating computers in constant time (or just in \sqrt{t} time)? The common wisdom is that the answer depends on the task in question: if a single person can dig at a rate of one cubic meter per hour, then in one hour a hundred people can dig a ditch that is 100 m long, but not a hole 100 m deep. Determining which computational tasks can be “parallelized” when many processors are available and which are “inherently sequential” is a basic question for both practical and theoretical reasons.

A very good feature of the circuit model is that it can easily be used to study questions of this kind. Let us define the *depth* of a DAG to be the length of the longest directed path in it: that is, the longest sequence of vertices where there is an edge from each one to the next. This notion of depth models the *parallel time* needed to compute the function: if you put a separate processor at each gate of a circuit of depth d , and at each phase you evaluate all gates for which the inputs have already been evaluated, then the number of phases you need is d . Parallel time is another important computational resource. Here again our knowledge is scarce—we do not know how to disprove the statement that every explicit function can be computed by a circuit of polynomial size *and* logarithmic depth.

Thus, we will restrict d to be a constant. It then becomes necessary to allow our gates to have *unbounded*

fan-in, meaning that the AND and OR gates are allowed to have any number of incoming edges. (If we do not allow this, then each output bit can depend only on a constant number of input bits.) With this very stringent restriction on circuit depth, it is possible to prove lower bounds for the complexity of explicit functions. For example, let $\text{PAR}(x)$ (for “parity”) equal 1 if and only if the binary string has an odd number of 1s, and let $\text{MAJ}(x)$ (for “majority”) equal 1 if and only if there are more 1s than 0s in x .

Theorem. *For any constant d , the functions PAR and MAJ cannot be computed by a polynomial-sized family of circuits of depth d .*

This result is due to another fundamental proof technique: the *random restriction method*. The idea is to fix at random (with judiciously chosen parameters) most of the input variables, by assigning them random values. Note that this simultaneously restricts the function as well as the circuit. This “restriction” should satisfy the following two properties.

- (i) The restricted circuit becomes very simple: for instance, it may depend on only a small subset of the remaining, unfixed input variables.
- (ii) The restricted function remains complex: for instance, it may depend on all remaining input variables.

For PAR the second property is easily seen to hold, and of course the heart of the matter is analyzing the effect of random restrictions on shallow circuits.

Interestingly, MAJ remains hard for constant-depth polynomial-size circuits even if the circuits are also allowed (unbounded fan-in) PAR-gates. However the “converse” does not hold; that is, PAR has constant-depth polynomial-size circuits with (unbounded fan-in) MAJ-gates. Indeed, the latter class seems to be quite powerful: nobody has managed to prove that there are functions in \mathcal{NP} that cannot be computed by such circuits, even if the depth is restricted to 3.

5.1.4 Formula Size

Formulas are perhaps the most standard way in which mathematicians express functions. For example, the largest root of the quadratic polynomial $at^2 + bt + c$, in terms of its (input) coefficients a , b , and c , is represented by the formula $(-b + \sqrt{b^2 - 4ac})/2a$. This is an arithmetic formula. In Boolean formulas the logical

operations “ \neg ”, “ \wedge ”, “ \vee ” replace the arithmetic operations above. For example, if $x = (x_1, x_2)$ is a Boolean string of length 2, then $\text{PAR}(x)$ is given by the formula $(\neg x_1 \wedge x_2) \vee (x_1 \wedge \neg x_2)$.

Any formula can be represented by a circuit, but this circuit has the additional property that its underlying DAG is a *tree*. Intuitively, this means that the computation is not allowed to reuse a previously computed partial result (unless it recomputes it). A natural size measure for formulas is the number of occurrences of variables in them, which is the same as the number of gates, to within a factor of 2.

Formulas are natural not only because of their prevalence in mathematics, but also because their size can be related to the depth of circuits and to the *memory* requirements of Turing machines (i.e., their space complexity).

By recursively using the above formula for PAR, that is, by using the fact that $\text{PAR}(x_1, \dots, x_{2n})$ is equal to $\text{PAR}(\text{PAR}(x_1, \dots, x_n), \text{PAR}(x_{n+1}, \dots, x_{2n}))$, we obtain a formula for the parity of n variables that has size n^2 . Given the fact that PAR has a simple circuit of linear size, one might wonder if there are smaller formulas as well. One of the oldest results in circuit complexity gives a negative answer.

Theorem. *Boolean formulas for PAR and MAJ must have at least quadratic size.*

The proof follows a simple combinatorial (or information-theoretic) argument. By contrast, there are linear-size circuits for both functions. This is very easy to show for PAR, but not for MAJ.

Can we give super-polynomial lower bounds on formula size? One of the cleanest methods suggested so far is the *communication complexity method*, which provides an information-theoretic setting for studying this computational problem. The power of this approach has been demonstrated mainly in the context of monotone formulas, where it yields an exponential lower bound for the PERFECT MATCHING problem (defined in section 5.1.2).

Suppose that two players play the following game. One player is given a graph G with n vertices that contains no perfect matching, and the other is given a graph H , with the same vertices, that does contain a perfect matching. Then there must be some pair of vertices that are joined by an edge in H but not joined in G . The aim of the two players is to find such a pair by sending each other bit strings, which each thinks of as encoding messages according to some prearranged

scheme. Of course, the player with graph G could simply send enough messages to specify the entire graph, but the question is whether there is some protocol that would enable them to find a pair of the desired kind with far fewer bits being exchanged. The smallest number of bits needed (in the worst case) is called the *monotone communication complexity* of the problem.

It has been shown that the monotone communication complexity must be at least linear in n , and this leads to the exponential lower bound just mentioned. More generally, if $f : \mathbb{I}_n \rightarrow \{0, 1\}$ is a monotone function, then the monotone communication complexity of f is the smallest number of bits that must be exchanged, in the worst case, to find a place i where $x_i = 0$ and $y_i = 1$, if $f(x) = 0$ and $f(y) = 1$. If f is not monotone, then one simply asks to find i such that x_i and y_i differ, and the smallest number of exchanges needed is the *communication complexity* of f . It can be shown that the monotone formula size of f is at least $\exp(cm)$ for a positive constant c if and only if the monotone communication complexity of f is at least $c'm$ for a positive constant c' . The corresponding statement also holds for general formula size and general communication complexity.

5.1.5 Why Is It so Difficult to Prove Lower Bounds?

We have seen that complexity theory has developed quite a few powerful techniques, which have at least been useful in proving strong lower bounds in restricted models of computation. But they all fall well short of providing nontrivial lower bounds for *general* circuits. Is there a fundamental reason for this failure? The same may be asked about any long-standing mathematical problem, such as THE RIEMANN HYPOTHESIS [V.29], for example, and the typical answer would be rather vague: that it seems that the current tools and ideas do not suffice.

Remarkably, for circuit complexity this vague feeling has been made into a precise theorem. Thus, there is a “formal excuse” for our failure so far. Roughly speaking, a very general class of arguments, called *natural proofs*, has been defined and shown to include all known proofs of lower bounds for restricted circuits. In fact, so broad is the class of arguments that it is very hard to envisage what an “unnatural” proof might be like. On the other hand, it has also been shown that if there is a natural proof that $\mathcal{P} \neq \mathcal{NP}$, then there are fairly efficient (not quite polynomial-time, but significantly faster than known) algorithms for various problems, including integer factorization. So if, like most

complexity theorists, you believe that these problems do *not* have efficient algorithms, then you also believe that *there is no natural proof that $\mathcal{P} \neq \mathcal{NP}$* .

The connection between natural proofs that $\mathcal{P} \neq \mathcal{NP}$ and some notoriously hard problems is through the notion of *pseudorandomness*, which is discussed in section 7.1.

One interpretation of this result is that it shows that general circuit lower bounds are “independent” of a certain natural fragment of PEANO ARITHMETIC [III.69]. This gives a hint that the \mathcal{P} versus \mathcal{NP} question may be independent of all of Peano arithmetic, or even of THE AXIOMS OF ZFC [IV.22 §3.1], although few believe the latter to be the case.

5.2 Arithmetic Circuits

As mentioned earlier, directed acyclic graphs can be used in various different contexts. We shall now leave Boolean functions and operations and look instead at arithmetical operations and functions that take numerical values, by which we mean values in \mathbb{Q} or \mathbb{R} or indeed in any FIELD [L3 §2.2]. If F is a field, then we can consider a DAG in which the inputs are now elements of F and the gates are the field operations “+” and “ \times ” (including multiplication by fixed field elements such as -1). Then, just as with Boolean circuits, once we know the inputs we can assign values to all vertices of the DAG: at each vertex one just applies the corresponding arithmetical operation to the values assigned to its predecessors, once these have been calculated. An arithmetic circuit computes a polynomial function $p : F^n \rightarrow F^m$, and every homogeneous polynomial function is computed by some circuit. To allow the computation of inhomogeneous polynomials, we augment the model by allowing a special input vertex whose value is the constant “1” of the field.

Let us consider a couple of examples. The polynomial $x^2 - y^2$, which as written requires two multiplications and one addition, can be computed by the circuit $(x + y)(x - y)$ which requires instead one multiplication and two additions. The polynomial x^d , which is defined using $d - 1$ multiplications, may in fact be computed with only $2 \log d$ multiplications: first compute x, x^2, x^4, \dots (each term in the sequence squaring the previous one), and then multiply together the appropriate subset of these powers to get the exponent d .

We denote by $S_F(p)$ the smallest possible size of a circuit that computes p . When we give no subscript, we shall assume that $F = \mathbb{Q}$, the field of rational numbers.

We do not count multiplication by a fixed field element as contributing to the size of a circuit: for example, when we said that $(x + y)(x - y)$ involves one multiplication, we were not counting the multiplication of y by -1 . The reader may wonder about division. However, we will be mainly interested in computing polynomials, and for computing polynomials (over infinite fields) division can be efficiently emulated by the other operations. As usual, we will be interested in sequences of polynomials, one for every input size, and will study size asymptotically.

It is easy to see that, for any *fixed* finite field F , arithmetic circuits over F can simulate Boolean circuits (on Boolean inputs) with only a constant factor increase in size. Thus, lower bounds for such arithmetic circuits yield corresponding lower bounds for Boolean circuits. Therefore, if we want to avoid the extreme difficulty with which we are already familiar, it makes sense to focus more on infinite fields, where lower bounds may perhaps be easier to obtain.

As in the Boolean case, the mere existence of hard polynomials is easy to establish.⁴ But, as before, we will be interested in *explicit* (families of) polynomials. The notion of explicitness is more delicate here, but it can be formally defined (and, for example, polynomials with algebraically independent coefficients are not considered explicit).

An important parameter, which is absent in the Boolean model, is the *degree* of the polynomial(s) being computed. For example, a polynomial of degree d , even in one variable, requires size at least $\log d$. Let us briefly consider the one-variable, or *univariate*, case first, in which the degree is the main parameter of interest, since this case already contains striking and important problems. Then we shall move to the general *multivariate* case, in which n , the number of inputs, will be the main parameter.

5.2.1 Univariate Polynomials

How tight is the $\log d$ lower bound for the size of an arithmetic circuit computing a polynomial of degree d ? A simple dimension argument shows that for most degree- d polynomials p , $S(p)$ is proportional to d . However, we know of no explicit polynomial with this

property. (Of course, this is shorthand for “explicit family of polynomials, one for each degree d .”) In fact, considerably less is known even than this.

Open problem. Find an explicit polynomial p of degree d , such that $S(p)$ is not bounded above by $c \log d$ for some constant c .

Two concrete examples are illuminating. Let $p_d(x) = x^d$, and $q_d(x) = (x + 1)(x + 2) \cdots (x + d)$. We have already seen that $S(p_d) \leq 2 \log d$, so the trivial lower bound is relatively tight. On the other hand, it is a major open problem to determine $S(q_d)$, and the conjecture is that $S(q_d)$ grows more quickly than any power of $\log d$. This question is particularly important because of the following result. *If $S(q_d)$ is bounded above by a power of $\log d$, then integer factorization has polynomial-size circuits.*

5.2.2 Multivariate Polynomials

Now let us return to polynomials with n variables. It is convenient to make n our only input size parameter, so we shall restrict ourselves to polynomials of total degree at most n , even when we do not mention this restriction.

For almost every polynomial p in n variables, $S(p)$ is at least $\exp(n/2)$. Again, this follows from an easy dimension argument, but again we would like to find explicit (families of) polynomials that are hard to compute. Unlike in the Boolean world, here there are lower bounds that slightly exceed the trivial ones. The following theorem is proved using elementary tools from algebraic geometry.

Theorem. There is a positive constant c such that $S(x_1^n + x_2^n + \cdots + x_n^n) \geq cn \log n$.

The same techniques extend to prove lower bounds of similar strength for other natural polynomials such as the symmetric polynomials and the DETERMINANT [III.15] (which can be regarded as a polynomial in the entries of the matrix). Establishing a stronger lower bound for some explicit polynomial is a major open problem. Another is obtaining a superlinear lower bound for any polynomial map of *constant total degree*. Outstanding candidates for the latter are the *linear* maps that compute the discrete Fourier transform over the complex numbers or the Walsh transform over the rationals. For both these transformations algorithms of time complexity $O(n \log n)$ are known.

4. A counting argument over infinite fields is inadequate (e.g., for every $a, b \in F$ the circuit $ax + b$ has size two, and so there are infinitely many circuits of size 2). Instead, a “dimension” argument is used, showing that the set of polynomials that are computable by small circuits forms a vector space of lower dimension than the set of all polynomials of adequate degree.

Now let us focus on specific polynomials of central importance. The most natural and well-studied candidate for the last open problem is MATRIX MULTIPLICATION [I.3 §4.2]: given two $m \times m$ matrices A, B , how many operations are needed to compute their product? The obvious algorithm, which follows from the definition of matrix product, requires about m^3 operations. Can this be beaten? It turns out that what really matters here is the number of multiplications. The first hint that one can improve on the obvious algorithm comes from the first nontrivial case (i.e., $m = 2$). While the usual algorithm uses eight multiplications, one can in fact reorganize the calculation and get away with only seven. This leads to a recursive argument: given a $2m \times 2m$ matrix, think of it as a 2×2 matrix, each entry of which is itself an $m \times m$ matrix. It follows that doubling the size of the matrix increases the number of multiplications needed by a factor of at most 7. This argument leads to an algorithm with only $m^{\log_2 7}$ multiplications (and roughly as many additions).

These ideas have been developed and extended to yield the following strong, but not quite linear, upper bound, where we denote by $n = m^2$ the natural input size, and by MM the matrix multiplication function.

Theorem. *For every field F there is a constant c such that $S_F(\text{MM}) \leq cn^{1.19}$.*

So what is the complexity of MM (even if one counts only multiplication gates)? Is it linear, or almost linear (something like $n \log n$, say), or is $S(\text{MM})$ at least n^α for some $\alpha > 1$? This is a famous open problem.

We next consider two polynomials in the $n = m^2$ variables representing an $m \times m$ matrix. We have already mentioned the determinant, but we shall also look at the *permanent*, which is defined by the determinant formula, except that now all the signs are positive. (In other words, one simply adds up $m!$ products instead of adding some and subtracting others.) We shall denote these by DET and PER, respectively.

While DET plays a major role in classical mathematics, PER is somewhat esoteric (though it appears in statistical mechanics and quantum mechanics). In the context of complexity theory both polynomials are of great importance, because they are representative of natural complexity classes. DET has relatively low complexity (and is related to the class of polynomials having polynomial-sized arithmetic formulas), while PER seems to have high complexity (indeed, it is complete for a complexity class of counting problems denoted

$\#P$, which extends \mathcal{NP}). Thus, it is natural to conjecture that PER is *not polynomial-time reducible* to DET.

One restricted type of reduction that makes sense in this algebraic context is called *projection*. Suppose we wish to find an algorithm for computing the permanent of an $m \times m$ matrix A . One approach might be to construct an $M \times M$ matrix B such that each of its entries is either a (variable) entry of A or a fixed element of the field, and to do so in such a way that the determinant of B equals the permanent of A . Then, as long as M is not too much larger than m , we can use the efficient algorithm for DET to give us an efficient algorithm for PER. A projection of this kind is known to exist with $M = 3^m$, but this is nothing like good enough. Therefore we ask the following question.

Open problem. *Can the permanent of an $m \times m$ matrix be expressed as the determinant of an $M \times M$ matrix, with M bounded above by a polynomial in m ?*

If so, then $P = \mathcal{NP}$: therefore, the answer is likely to be no. Conversely, if the answer could be shown to be no, then this would provide a significant step toward proving that $P \neq \mathcal{NP}$, though it would probably not imply it.

5.3 Proof Complexity

The concept of *proof* distinguishes mathematics from all other fields of human inquiry. Mathematicians have gathered millennia of experience to attribute such adjectives to proofs as “insightful,” “original,” “deep,” and, most notably, “difficult.” Can one quantify mathematically the difficulty of proving various theorems? This is exactly the task undertaken in proof complexity. It seeks to classify theorems according to the difficulty of proving them, much as circuit complexity seeks to classify functions according to the difficulty of computing them. In proofs, just as in computation, there will be a number of models, called *proof systems*, that capture the power of reasoning that is allowed to the prover.

The types of statements, theorems, and proofs we shall deal with are best illustrated by the following example. We warn the reader in advance that the theorem we are about to discuss may seem too trivial to give us any insight into the nature of proofs: however, it turns out to be highly relevant.

The theorem in question is the well-known *pigeon-hole principle*, which states that if you have more

pigeons than holes then at least two pigeons will have to share a hole. More formally, there is no INJECTION [I.2 §2.2] f from a finite set X to a smaller finite set Y . Let us reformulate this theorem and then discuss the complexity of proving it. First, we turn it into a sequence of finite statements. For each $m > n$ let PHP_n^m stand for the statement, “You cannot fit m pigeons into n holes if each pigeon needs a hole to itself.” A convenient way of formulating this mathematically is to use an $m \times n$ matrix of Boolean variables x_{ij} . This can be used to describe a hypothetical mapping if we interpret $x_{ij} = 1$ to mean that the i th pigeon is placed in the j th hole. The pigeonhole principle states that either some pigeon is not mapped anywhere or two pigeons are mapped to the same hole. In terms of the matrix, this says that either there is some i such that $x_{ij} = 0$ for every j , or we can find $i \neq i'$ and j such that $x_{ij} = x_{i'j} = 1$.⁵ These conditions are easily expressible as a *propositional formula* in the variables x_{ij} (that is, an expression built out of the x_{ij} using “ \wedge ”, “ \vee ”, and “ \neg ”), and the pigeonhole principle is the statement that this formula is a *tautology*: that is, it is satisfied by *every* assignment of true or false values (or equivalently 1 or 0) to the variables.

How can we prove this tautology to someone who can read our proof and perform simple, efficient computations? Here are a few possibilities which differ from each other in a number of ways.

- The standard proof uses symmetry and induction. It reduces PHP_n^m to PHP_{n-1}^{m-1} by saying that once the first pigeon has been assigned a hole, the task that is left is to place the remaining $n - 1$ pigeons into $m - 1$ holes. Notice that these holes may not be the *first* $n - 1$ holes, so for such an argument to become a formal proof one must argue by symmetry. Our proof system must be strong enough to capture this symmetry (which amounts to a renaming of the variables), and it must also allow us to use induction.
- At the other extreme, one can obtain a trivial proof, which requires only “mechanical reasoning,” by simply presenting an evaluation of the formula for every possible input. As there are mn variables, the proof length is 2^{mn} , which is exponential in the size of the formula describing the assertion PHP_n^m .

5. Note that we have not ruled out the possibility that some pigeon is mapped to more than one hole—we could do so, but the principle remains valid even if we do not.

- A more sophisticated (“mechanical”) proof uses counting. Assume for a contradiction that there exists an assignment of truth values to the variables that falsifies the formula. Since each pigeon is mapped to some hole, the assignment must have at least m 1s. But since each hole contains at most one pigeon, the assignment must contain at most n 1s. Therefore, $m \leq n$, which contradicts the assumption that $m > n$. For this proof to be admissible, our system has to allow inferences powerful enough to do counting of this kind.

The lesson from the above example is that proofs and their length depend on the underlying proof system. But what exactly is a proof system, and how do we measure the complexity of a proof? It is to this question that we now turn. Here are the salient features that we expect from any such system.

Completeness: every true statement has a proof.

Soundness: no false statement has a proof.

Verification efficiency: given a mathematical statement T and a purported proof for it π , it can be easily checked whether π does indeed prove T in the system.⁶

Actually, even the first two requirements are too much to expect from strong proof systems, as GÖDEL [VI.92] famously proved in his INCOMPLETENESS THEOREM [V.18]. However, we are considering just propositional formulas with finite proofs, and for these there are proof systems. In this context, the above conditions are concisely captured by the following definition.

Definition. A (*propositional*) *proof system* is a polynomial-time Turing machine M with the property that T is a tautology if and only if there exists a (“*proof*”) π such that $M(\pi, T) = 1$.⁷

As a simple example, consider the following “truth-table” proof system M_{TT} , which corresponds to the trivial proof in the foregoing example. Basically, this machine will declare a formula T to be a theorem if evaluating T on each possible input makes T true. A bit more formally, for any formula T in n variables, $M_{\text{TT}}(\pi, T) = 1$ if and only if π is a list of all binary

6. Here, efficiency of the verification procedure refers to its running time measured in terms of the *total length of the alleged theorem and proof*. In contrast, in sections 3.2 and 6.3, we consider the running time as a function of the *length of the alleged theorem* (or, alternatively, allow only proofs of a priori bounded length).

7. In agreement with standard formalisms (see below), the proof is seen as coming before the theorem.

strings of length n , and for each such string σ we have $T(\sigma) = 1$.

Notice that M_{TT} runs in polynomial time in its input length. The point, of course, is that for typical interesting formulas such as the pigeonhole principle, whose size depends polynomially on the number of variables, the input length is extremely long, since the proof π has length exponential in the length of the formula. This leads us to the definition of the efficiency (or complexity) of a general propositional proof system M : it is the length of the shortest proof of each tautology. That is, if T is a tautology, we define its complexity $\mathcal{L}_M(T)$ to be the length of the shortest string π such that $M(\pi, T) = 1$. We then measure the efficiency of the proof system itself (i.e., M) by defining $\mathcal{L}_M(n)$ to be the maximum of $\mathcal{L}_M(T)$ over all tautologies T of length n .

Is there a propositional proof system which has polynomial-size proofs for all tautologies? The following theorem provides a basic connection between this question and computational complexity, and in particular with the major question of section 3.4. It follows quite easily from the NP-completeness of SAT, the problem of satisfying propositional formulas (and the fact that a formula is satisfiable if and only if its negation is not a tautology).

Theorem. *There exists a proof system M such that \mathcal{L}_M is polynomial if and only if $\mathcal{NP} = \text{co } \mathcal{NP}$.*

To start attacking this formidable problem it makes sense to begin by considering simpler (and thus weaker) proof systems, before moving on to more and more complex ones. Moreover, there are tautologies and proof systems that naturally suggest themselves as good ones to study, systems in which certain basic forms of reasoning are allowed while others are not. In the rest of this section we shall focus on some of these restricted proof systems.

If a typical proof in a branch of mathematics such as algebra, geometry, or logic is written out in full, then it starts with some axioms and proceeds to a conclusion using a set of very simple and transparent *deduction rules*. Each line of the proof consists of a mathematical statement, or formula, which follows from earlier statements by means of one of these rules.⁸ This deductive approach goes right back to EUCLID [VI.2] and perfectly fits our DAG model: the inputs can be labeled

by the axioms, every other vertex is assigned a deduction rule, and the statement associated with each vertex is the statement that follows from its predecessors by means of the specified rule.

There is an equivalent and somewhat more convenient view of (simple) proof systems, namely as (simple) *refutation* systems. These encapsulate the idea of a proof by contradiction. We assume the negation of the tautology T we wish to prove, and use the rules of the system to derive a contradiction—that is, a statement that is identically FALSE. It is often easy to write the negation of a tautology T as a conjunction of mutually contradicting formulas (e.g., a set of clauses with no common truth assignment, a system of polynomials with no common root, a collection of half-spaces with empty intersection, etc). Assuming, for a contradiction, that all these *are* simultaneously satisfiable by some σ (which could be an assignment, root, or point, respectively), we derive more and more formulas that must also be satisfied by σ because of the soundness of the derivation rules, until eventually we reach a blatant contradiction (such as $\neg x \wedge x$, $1 = 0$, or $1 < 0$, respectively). We will use the refutation viewpoint throughout, and often exchange “tautology” and its negation, “contradiction.”

So we turn to studying the proof length $\mathcal{L}_\Pi(T)$ of tautologies T in proof systems Π . The first observation, which reveals a major difference between proof complexity and circuit complexity, is that the trivial counting argument *fails*. The reason is that, while the number of functions on n bits is 2^{2^n} , there are at most 2^n tautologies of length n . Thus, in proof complexity, even the *existence* of a hard tautology, let alone an explicit one, would be of interest. As we shall see, however, most known lower bounds (in restricted proof systems) apply to very natural tautologies.

5.3.1 Logical Proof Systems

The proof systems in this section will all have lines that are Boolean formulas. The differences between the systems will be in the structural limits that are imposed on these formulas.

The most basic proof system, called the *Frege system*, puts no restriction on the formulas manipulated by the proof. It has just one derivation rule, called the *cut rule*: from the two formulas $(A \vee C)$, $(B \vee \neg C)$ we can derive $A \vee B$. Different basic books in logic have slightly different ways of describing this system. However, from a computational perspective they are all equivalent, in

8. General proof systems as we defined them can also be adapted to this formalism, by considering a deduction rule that corresponds to a single step of the machine M . However, the deduction rules considered below are even simpler, and more importantly they are natural.

the sense that (up to polynomial factors) the length of the shortest proofs is independent of which variant you pick.

The counting-based proof of the pigeonhole principle can be carried out efficiently in the Frege system (but this is not a trivial fact), which tells us that $\mathcal{L}_{\text{Frege}}(\text{PHP}_n^{n+1})$ is polynomial in n . The major open problem in proof complexity is to find any tautology (as usual we mean a family of tautologies) that has no polynomial-size proof in the Frege system.

Open problem. *Establish super-polynomial lower bounds for the Frege system.*

As it seems to be very hard to find lower bounds for Frege systems, we turn to natural and interesting subsystems. The most widely studied system is called *resolution*. Its importance stems from its use by most propositional (as well as first-order) *automated theorem provers*.⁹ The formulas allowed in resolution refutations are simply clauses (disjunctions), so the cut rule defined earlier simplifies to the *resolution rule*: from two clauses $(A \vee x)$, $(B \vee \neg x)$ we can derive $A \vee B$, where A, B are clauses and x is a variable.

A major result of proof complexity is that proving the pigeonhole principle is hard in the resolution system.

Theorem. $\mathcal{L}_{\text{resolution}}(\text{PHP}_n^{n+1}) = 2^{\Omega(n)}$

The proof of this result is related in an interesting way to the circuit lower bounds for the parity and majority functions discussed in section 5.1.3.

5.3.2 Algebraic Proof Systems

Just as a natural contradiction in the Boolean setting is an unsatisfiable collection of clauses, a natural contradiction in the algebraic setting is a system of polynomials without a common root.¹⁰

How would you prove that the system $\{f_1 = x\mathcal{Y} + 1, f_2 = 2yz - 1, f_3 = xz + 1, f_4 = x + y + z - 1\}$ has no common root (over any field)? A quick way is to

observe that $zf_1 - xf_2 + yf_3 - f_4 \equiv 1$. Clearly, a common root of the system would be a root of this linear combination, which is a contradiction because the constant 1 function has no root. Can we always use such proofs?

A famous theorem, HILBERT'S NULLSTELLENSATZ [V.20], tells us that the answer is yes. It states that if f_1, f_2, \dots, f_n are polynomials (with any number of variables) that have no common root, then there exist polynomials g_1, \dots, g_n such that $\sum_i g_i f_i \equiv 1$. How efficient are such proofs? Can we always have proofs (i.e., g_i s) of length polynomial in the description of the f_i s? Unfortunately not: the shortest explicit description of the g_i s may be of exponential length, though proving this fact is highly nontrivial.

Another natural proof system, which is related both to Hilbert's Nullstellensatz and to computations of Gröbner bases in symbolic algebra programs, is *polynomial calculus* (PC). The lines in this system are polynomials, represented explicitly by all their coefficients, and it has two deduction rules: for any two polynomials g, h , we can derive their sum, $g + h$, and for any polynomial g and any variable x_i , we can derive the product $x_i g$. PC is known to be exponentially stronger than the proof system underlying Hilbert's Nullstellensatz. However, strong size lower bounds (obtained from degree lower bounds) are known for this system as well. For example, encoding the pigeonhole principle as a contradicting set of constant degree polynomials, we have the following theorem.

Theorem. *For every n and every $m > n$, $\mathcal{L}_{\text{PC}}(\text{PHP}_n^m) \geq 2^{n/2}$, over every field.*

5.3.3 Geometric Proof Systems

Yet another natural way to represent contradictions is by sets of regions in space that have empty intersection. For instance, many important problems in *combinatorial optimization* concern systems of linear inequalities in \mathbb{R}^n and their relationship to the Boolean cube $\{0, 1\}^n$. Each inequality defines a half-space, and the problem is to decide whether the intersection of all these half-spaces contains a point with coordinates all equal to 0 or 1.

The most basic proof system is called *Cutting Planes* (CP). A line of a proof is a linear inequality with integer coefficients. The deduction rules are that you can add two inequalities, and, less obviously, that you can divide the coefficients by a constant and do some

PUP: Tim thinks the footnote called here is relevant. OK?

9. These are algorithms that attempt to generate proofs for given tautologies. These tautologies may be boring mathematically but of great practical importance, such as the statement that a computer chip or communication protocol functions correctly. Interestingly, popular applications also include a variety of theorems that are mathematically interesting, such as results in basic number theory.

10. Moreover, polynomials can easily encode propositional formulas. First, one puts such a formula into *conjunctive normal form*, or CNF: that is, one expresses it as the conjunction of a collection of clauses. CNF formulas can easily be converted to a system of polynomials, one per clause, over any field. One often adds the polynomials $x_i^2 - x_i$, which ensure Boolean values.

rounding, taking advantage of the fact that the points of the solution space have integer coordinates.

While PHP_n^m is easy in this system, exponential lower bounds are known for other tautologies. They are obtained from the monotone circuit lower bounds of section 5.1.2.

6 Randomized Computation

Up to now, the computations we have considered have all been *deterministic*: that is, the output is completely determined by the inputs and the rules governing the computations. In this section we shall continue to focus on polynomial-time computations, but now we shall allow our computing devices to make *probabilistic*, or *randomized*, choices.

6.1 Randomized Algorithms

A famous example of such an algorithm is one that tests for primality. If N is the positive integer to be tested, then the algorithm randomly chooses k numbers less than N , and repeatedly performs a simple test using each of the chosen numbers in turn. If N is composite, then the probability that the test detects this is at least $\frac{1}{4}$. Therefore, the probability that the algorithm fails to detect it for any of the k numbers is at most $(\frac{3}{4})^k$, which is very small indeed for even modestly large values of k . Details of how the test works can be found in COMPUTATIONAL NUMBER THEORY [IV.3 §2].

It is not hard to give a rigorous definition of a randomized Turing machine, but we shall not need the precise details here. The main point is that if M is a randomized Turing machine and x is an input string, then $M(x)$ is not a fixed output string, but rather a RANDOM VARIABLE [III.73 §4]. If, for example, the output is a single bit, then we shall make statements such as, “The probability that $M(x) = 1$ is p .” The actual value of $M(x)$ will depend on the particular random choices made by the machine M when it runs.

If we are using a randomized algorithm to solve a decision problem S , then we would like $M(x)$ to give the correct answer with high probability *whatever the input* x . (The correct answer is 1 if $x \in S$ and 0 otherwise.) This leads to the definition of the complexity class \mathcal{BPP} (for *bounded error, probabilistic polynomial time*).

Definition (\mathcal{BPP}). A Boolean function f is in \mathcal{BPP} if there exists a probabilistic polynomial-time machine M such that $\Pr[M(x) \neq f(x)] \leq \frac{1}{3}$ for every $x \in \mathcal{I}$.

The error bound $\frac{1}{3}$ is arbitrary, and can be made much smaller if one runs the algorithm several times and takes a majority vote of the answers. (We stress that the random moves in the various runs are independent.) Standard probabilistic estimates show that, for any k , the error probability can be reduced to 2^{-k} if one runs the algorithm $O(k)$ times.

Because randomness is believed to be “available” and an exponentially small chance of failure is of no practical importance, the class \mathcal{BPP} is in many ways a better model for efficient computation than \mathcal{P} , which it trivially contains. Let us mention some relations of this class \mathcal{BPP} to other complexity classes we have seen already. It is easy to see that $\mathcal{BPP} \subseteq \mathcal{EXP}$; if the machine tosses m coins, we could enumerate all 2^m possible outcomes of these coin tosses and take a majority vote. The relation of \mathcal{BPP} to \mathcal{NP} is not known, but it is known that if $\mathcal{P} = \mathcal{NP}$ then $\mathcal{P} = \mathcal{BPP}$ as well. Finally, nonuniformity can replace randomness: every function in \mathcal{BPP} has polynomial-size circuits. But the fundamental question is whether or not randomized algorithms are genuinely more powerful than deterministic ones (for decision problems).

Open problem. Does $\mathcal{P} = \mathcal{BPP}$?

As we mentioned earlier, a deterministic polynomial-time algorithm was recently discovered for primality testing, though in practice the randomized algorithm is much more efficient. However, there are quite a few problems¹¹ that are known to be in \mathcal{BPP} but not known to be in \mathcal{P} . Indeed, for most of these problems randomness gives an exponential improvement over the best deterministic algorithms that are known. Is this evidence that randomness increases our power to solve decision problems? Surprisingly, a completely different kind of evidence (discussed in section 7.1) suggests the opposite, namely that $\mathcal{P} = \mathcal{BPP}$.

6.2 Counting at Random

One important general question regarding \mathcal{NP} search problems is that of determining *how many* solutions a particular instance has. This includes a host of interesting problems from various disciplines: for example, counting the number of solutions to a system of multivariate polynomials, counting the number of perfect matchings of a graph (or, equivalently, computing the permanent of a $\{0, 1\}$ matrix), computing the volume

11. A central example is Identity Testing: given an arithmetic circuit over \mathcal{Q} , decide if it computes the identically zero polynomial.

of a polytope (defined by linear inequalities) in high dimension (see [I.4 §9] for more about this problem), computing various parameters of physical systems, etc.

For most of these problems, even approximate counting is good enough. Clearly, an approximate count of the number of solutions will in particular allow one to determine whether a solution exists at all. For example, if one knows the approximate number of satisfying assignments for a given propositional formula, then one certainly knows whether this number is at least 1. This tells us whether the formula is satisfiable and solves an instance of SAT. Interestingly, the converse is also true: if one can solve SAT, then one can use this ability to produce a randomized algorithm for approximating the *number* of solutions, to within any constant factor greater than 1. More precisely, there is an efficient probabilistic algorithm that can produce such an approximate count if it is allowed to make free use of a subroutine that solves SAT instances. It turns out that analogous statements holds for all NP-complete problems.

For some problems, approximate counting can be done *without* the SAT subroutine. There are polynomial-time probabilistic algorithms for approximating the permanent of positive matrices, approximating the volume of polytopes, and more. These algorithms use a connection between approximate counting and another natural algorithmic problem: that of randomly generating a solution in such a way that all correct solutions are equally likely to occur. The basic technique is to construct a Markov chain on the space of solutions with uniform stationary distribution and to analyze the rate of convergence of the chain to this distribution (see Hochbaum 1996, chapter 12).

What about *exact* counting? It is believed that this *cannot* be done by an efficient probabilistic algorithm, even if it can make free use of a SAT subroutine. A remarkable “complete” problem for this class of counting problems is counting the number of perfect matchings in a graph. What is surprising about it is that there is an efficient algorithm for finding a perfect matching in a graph, if one exists, and yet counting such matchings is complete in the sense that an efficient algorithm for doing this can be turned into an efficient algorithm for the counting version of *any other* problem in \mathcal{NP} .

6.3 Probabilistic Proof Systems

As we saw earlier, proof systems are defined in terms of their verification procedure. In section 5.3, we considered verification procedures that run in time that

is polynomial in the combined length of the assertion and its alleged proof. Here (as in section 3.2), we restrict our attention to verification procedures that run in time that is polynomial in the length of the assertion. Such proof systems are related to the class \mathcal{NP} , since sets S in \mathcal{NP} are those with the following property: there is a polynomial-time algorithm M such that x belongs to S if and only if there exists a string y of length polynomial in x with $M(x, y) = 1$. In other words, we can regard y as a concise proof (verifiable by M) that x belongs to S .

What if we now allow M to be a *randomized* algorithm? Then we obtain a *probabilistic proof system*. Such systems are not put forward as a substitute for the notion of mathematical proof, but rather as an interesting extension of the notion of efficient verifiability in situations where a tiny amount of error can be tolerated. As we shall see, various types of probabilistic proof systems yield enormous advantages in computer science. We shall exhibit three remarkable manifestations of this. The first shows that we can use it to prove many more theorems, the second that we can do so without revealing *anything* in our proof, and the third that alleged proofs can be written in such a way that verifiers need only look at a tiny handful of bits in order to decide whether they are correct.

6.3.1 Interactive Proof Systems

Recall the graph isomorphism problem from section 4. Given two graphs G and H , it asks whether H is obtained from G by simply permuting the vertices. This problem is clearly in \mathcal{NP} , since one can just exhibit a permutation that transforms G into H .

We can look at this as a protocol involving a verifier, who can do polynomial-time computations, and a prover, who has unlimited computational resources. The verifier wishes to be convinced that G and H are isomorphic, so the prover sends a permutation and the verifier checks (in polynomial time) that it is valid.

Suppose that we now look at the graph *nonisomorphism* problem. Is there any way for a prover to convince a verifier that two graphs G and H are not isomorphic? Obviously there will be for some pairs of graphs (G, H) , but there does not seem to be a systematic method of demonstration that works for *all* nonisomorphic pairs. Yet, remarkably, if we allow *random-*

ness and interaction, then there is a simple way for the verifier to be convinced.¹²

Here is how it works. The verifier chooses at random one of the two graphs G and H , randomly permutes its vertices, and sends it to the prover. The prover then sends back a message saying whether this permuted graph is G or H .

If G and H are not isomorphic, then the permuted graph is isomorphic to exactly one of G and H , so the prover can determine which and thereby get the right answer. But if G and H are isomorphic, then the prover has no way of knowing which graph has been permuted, and therefore has a 50% chance of getting the right answer.

So now, to become convinced, the verifier repeats the procedure k times. If the graphs are not isomorphic, the prover will always get the right answer. If they are isomorphic, then with probability $1 - 2^{-k}$ the prover will make at least one mistake. If k is large, this becomes a near-certainty, so if the prover never makes a mistake, then the verifier will be convinced that the graphs are not isomorphic.

That was an example of an *interactive proof system*. Given a decision problem S , an interactive proof system for S is a protocol involving an interacting verifier and prover, with the property that if $x \in S$ then the verifier will eventually output 1, while if $x \notin S$ then there is a probability of at least $\frac{1}{2}$ that the verifier will output 0. As in the example, the verifier can then repeat the protocol several times, thereby replacing $\frac{1}{2}$ by a probability very close to 1. Also as in the example, the verifier is allowed polynomial-time randomized computations and the prover has unlimited computational power. Finally, the number of rounds of the interaction must be at most polynomial in the size of the input x , so that the entire verification procedure is efficient. The class of decision problems for which an interactive proof system exists is denoted \mathcal{IP} .

One can view the protocol as an “interrogation” by a persistent student, who asks the teacher “tough” questions in order to be convinced of correctness. Interestingly, it turns out that asking “tough” questions is no better than asking random questions! That is, every set that has an interactive proof system also has one in which the verifier only asks random questions that are uniformly and independently distributed in some predetermined set.

12. We note that allowing interaction without randomness does not yield any gain; that is, such interactive (but deterministic) proof systems are exactly as powerful as \mathcal{NP} .

It turns out that for *every* decision problem S that belongs to \mathcal{NP} there is an interactive proof system that can be used to demonstrate that $x \notin S$. It works by demonstrating the *nonexistence of an NP-proof that x is in S* . The proof of this result, which tells us that $\text{co}\mathcal{NP} \subset \mathcal{IP}$, involves an arithmetization of Boolean formulas. Furthermore, a complete characterization of the power of interactive proofs is known. Let \mathcal{PSPACE} be the class of all problems solvable in polynomial *space* (or *memory*). Although solving problems in \mathcal{PSPACE} may require *exponential* time, they all have interactive proofs.

Theorem. $\mathcal{IP} = \mathcal{PSPACE}$.

While it is not known if $\mathcal{NP} \neq \mathcal{PSPACE}$, it is widely believed to be the case, and so it seems that interactive proofs are much more powerful than standard noninteractive and deterministic proofs (that is, NP-proofs).

6.3.2 Zero-Knowledge Proof Systems

A typical mathematical proof not only guarantees the truth of a statement, but also *teaches* you something about it. In this section we shall discuss a kind of proof that teaches you absolutely nothing, beyond the fact that the statement is true. Since this seems impossible, let us give an example.

Suppose a prover wants to convince you that a certain map (in the geography sense) can be colored with three colors in such a way that no two adjacent regions have the same color. The most obvious approach is actually to show you a coloring, but this teaches you something—a particular coloring—which you would not otherwise be able to find easily, even knowing that it existed (since this search problem is NP-complete). Is there any way the prover can convince you without giving you this extra knowledge?

Here is a way of doing it. Given any coloring of the map, with red, blue, and green, say, one can produce other colorings by permuting the colors: for instance, one might change all the red regions into blue and all the blue ones into red. Let the prover take six copies of the map and color them in six different ways, one for each permutation of the three colors. Now we have a sequence of rounds. In each round the prover randomly chooses one of the six colored maps, you randomly choose a pair of adjacent regions, and the prover allows you to check that they have different colors, *but does not allow you to look at the rest of the map*. If the graph

cannot be properly colored with three colors and the prover tries to cheat, then after enough rounds (a polynomial number suffices) you will discover the deception by hitting upon two adjacent regions that have been given the same color (or perhaps one of them has not been colored at all). However, at each stage, all you learn about the two regions you look at is that they have different colors—you have no idea what those colors are in the coloring the prover started with. So you end up with no knowledge beyond the fact that the map can (almost certainly) be properly colored.

Similarly, a “zero-knowledge proof” that a certain formula is satisfiable should not reveal a satisfying assignment, or even any partial information (such as the truth value of one of the variables), or irrelevant information that is hard to compute (such as how to factorize an integer that happens to be encoded by the formula). In general, a zero-knowledge proof is an interactive proof that does not help you (the verifier) to make *any* computations that you were not able to make efficiently already.

Which theorems have zero-knowledge proofs? Obviously, if the verifier can determine the answer with no help, then the theorem has a trivial zero-knowledge proof, in which the prover does nothing at all. Thus, any set in BPP has a zero-knowledge proof. The zero-knowledge proof outlined for 3-colorability depended on noncomputational procedures, such as the prover watching carefully to make sure that you just look at two regions. Implementing the protocol in full on a computer takes some care, but a method of doing it has been devised, which depends on the hardness of integer factorization. The result is a *zero-knowledge proof system*. Combining this with the NP-completeness of 3-colorability, one can prove that zero-knowledge proof systems exist for *every* set in NP . More generally, we have the following theorem.

Theorem. *If one-way functions exist (these are defined in section 7), then every set in NP has a zero-knowledge proof system. Moreover, this proof system can be efficiently derived from the standard NP proof.*

This theorem has a dramatic effect on the design of cryptographic protocols (see section 7.2). Furthermore, under the same assumption, an even stronger result holds: any set that has an interactive proof system also has a zero-knowledge interactive proof system.

6.3.3 Probabilistically Checkable Proofs

In this section we turn to one of the deepest and most surprising discoveries about the power of probabilistic proofs. Here, as in the case of standard (noninteractive) proofs, the verifier receives a complete written proof. The catch is that the verifier may read only a very small, randomly selected, part of this proof.

A good analogy is to imagine that you are refereeing a paper and trying to decide the correctness of a long proof by reading just a few random lines. If the proof has a single (but crucial) mistake, then you will probably not read the relevant line so you will not notice the mistake. But this is true only for the “natural” way of writing down proofs. It turns out that there are ways of writing proofs “robustly” (with a certain amount of redundancy) so that any mistake will manifest itself in many different places. (This may remind you of ERROR-CORRECTING CODES [VII.6]. There is indeed an important analogy here, and cross-fertilization between the two areas has been very significant.) Such a robust proof system is called a PCP, which stands for “probabilistically checkable proof.”

Loosely speaking, a *PCP system* for a set S consists of a probabilistic polynomial-time verifier who has access to individual bits in a string that represents the (alleged) proof. The verifier tosses coins and, depending on the outcome, accesses only a *constant* number of the bits in the alleged proof. It should output 1 whenever x belongs to S (and an adequate proof is provided), while if x does not belong to S , then (no matter which false proof is provided) it should output 0 with probability at least $\frac{1}{2}$.

Theorem (the PCP theorem). *Every set in NP has a PCP system. Furthermore, there exists a polynomial-time procedure for converting any NP-proof to the corresponding PCP.*

In particular, it follows that the (robust) PCP has length that is polynomial in the length of the input. In fact, this PCP is itself an NP-proof.¹³

On top of its direct conceptual appeal, the PCP theorem (and its variants) has a major application to complexity theory: it allows us to prove that several natural approximation problems are hard (assuming that $P \neq NP$).

13. Here we take advantage of the fact that PCP systems are defined to be error free when $x \in S$ and the fact that the verifier in the PCP theorem uses only a logarithmic number of coin tosses, so one can efficiently check all possible outcomes.

For example, suppose we are given n linear equations over the two-element field \mathbb{F}_2 . If we choose random values for the variables, then any given equation will be satisfied with probability $\frac{1}{2}$, so it is clearly possible to satisfy at least half the equations. Also, by linear algebra one can quickly determine whether it is possible to satisfy all the equations simultaneously. However, it turns out that if $\mathcal{P} \neq \mathcal{NP}$ then there is no polynomial-time algorithm that will output 1 if 99% of the equations can be satisfied simultaneously and 0 if it is impossible to satisfy more than 51% of them. That is, even *approximately* determining the number of equations that can be satisfied simultaneously is hard.

To see the connection between such approximation problems and PCP, note that a PCP system for any set S gives rise to an optimization problem as follows. Suppose we are given an input x . Then for any alleged proof that $x \in S$, which is presented as a string y , there is a certain probability that the verifier accepts y . What is the maximum of this probability over all alleged proofs y ? If we could answer this question to within a factor of 2, then we would be able to tell whether x belongs to S . Hence, if S is an NP-complete decision problem, the PCP theorem implies that this optimization problem is NP-hard (that is, at least as hard as any problem in \mathcal{NP}). One can now use reductions, capitalizing on the fact that the verifier reads only a constant number of bits in the alleged proof, to obtain similar results for many natural optimization problems.

This is of great theoretical interest, but some practical disappointment: in many cases, approximate solutions would have been just as useful as exact ones, but they turn out to be just as hard to obtain.

6.4 Weak Random Sources

We now turn to the question of how to obtain the randomness for all the probabilistic computations discussed in this section. Although randomness seems to be present in the world (e.g., the perceived randomness in the weather, Geiger counters, Zener diodes, real coin flips, etc.), it does not seem to be in the perfect form of the unbiased and independent coin tosses we have postulated. If we actually want to use randomized procedures, then we need to convert weak sources of randomness into almost perfect ones, because this is what probabilistic computations were defined to work with.

Algorithms that convert imperfect randomness into a stream of almost completely independent and unbiased bits are called *randomness extractors*, and near

optimal ones have been constructed. This large body of work is surveyed in Shaltiel (2002), for example. The questions that arise turn out to be related to certain types of pseudorandom generators (see section 7.1) as well as to combinatorics and coding theory.

To illustrate the nature of the problem of randomness extraction, we consider three relatively simple models of weak random sources. Imagine first that you are in possession of a biased coin that has probability p of coming up Heads, where $\frac{1}{3} < p < \frac{2}{3}$, but you do not know the bias. Can you produce a uniformly distributed binary value using such a coin? A simple solution consists of tossing the coin twice, outputting 1 if the result is Heads followed by Tails and 0 if the result is Tails followed by Heads, and otherwise continuing to the next attempt. This way we can generate a perfect coin toss by tossing the biased coin an expected number $((1-p)p)^{-1}$ of times.

A more challenging setting arises if you are given n different biased coins, with unknown biases p_1, \dots, p_n , each in the interval $(\frac{1}{3}, \frac{2}{3})$, and you are asked to generate an almost uniformly distributed binary value by tossing each of these coins *exactly once*. Here a good solution consists of tossing all coins and outputting the parity of the number of Heads. It can be shown that the outcome will be 1 with a probability that is exponentially (in n) close to $\frac{1}{2}$.

Finally, consider a situation in which the devil designs the coins in the latter example, but does so after seeing the outcome of previous coin tosses. That is, you are tossing n different coins, but the bias of the i th coin (i.e., p_i) may depend on the outcome of the previous $i-1$ coin tosses (but still lies between $\frac{1}{3}$ and $\frac{2}{3}$). It can be shown that in this case you cannot do better than simply outputting the outcome of the first coin. However, if you are allowed to use just a few genuinely random bits, then you can do much better: given just $O(\log(n/\epsilon))$ perfectly random coin tosses, together with the n biased coin tosses, you can output a string of length proportional to n that is “ ϵ -close” to being uniformly distributed.

7 The Bright Side of Hardness

If $\mathcal{P} \neq \mathcal{NP}$, as almost everybody believes, then there are computational problems of great interest that are inherently intractable. This is bad news, but there is a bright side to the matter: computational hardness has many fascinating conceptual consequences as well as important practical applications.

The hardness assumption we shall make is the existence of *one-way functions*; namely, functions that are easy to compute but hard to invert. For example, the product of two integers is of course easy to compute, but its “inverse”—factoring the resulting product—is the integer factorization problem, widely believed to be intractable. For our purposes, we shall need the inverse to be hard not just in the worst case, but hard *on average*. For example, for factoring it is believed that the product of two random primes of length n cannot be factored in polynomial time, even with some small constant probability of success. In general, we shall say that a function $f : \mathbb{I}_n \rightarrow \mathbb{I}_n$ is a *one-way function* if it is easy to evaluate (i.e., there exists a polynomial-time algorithm that returns $f(x)$ when you input x) but hard to invert in the following average-case sense: any polynomial-time algorithm M will fail to invert f correctly for at least half the input strings $x \in \mathbb{I}_n$. That is, for at least half the strings x , if you input $y = f(x)$ into M , then the output will not be a string x' such that $f(x') = y$.

Do one-way functions exist? It is easy to see that if $\mathcal{P} = \mathcal{NP}$ then the answer is no. The converse is an important open problem: *If $\mathcal{P} \neq \mathcal{NP}$, does it follow that one-way functions exist?*

Below, we discuss the connections between computational difficulty (in the form of one-way functions), and two important computational complexity theories: the theory of *pseudorandomness* and the theory of *cryptology*.

7.1 Pseudorandomness

What is randomness? When should we say that a mathematical or physical object behaves randomly? These are fundamental questions that have been thought about for centuries. When the objects are probability distributions, on n -bit sequences, say, there is consensus about one point at least: the *uniform* distribution (in which each n -bit string appears with probability 2^{-n}) is “the most random” one. More generally, it seems reasonable to say that any distribution that is statistically close to the uniform distribution should also be regarded as having “good randomness” properties.¹⁴

One of the great insights of computational complexity theory is that there are distributions that are extremely far from the uniform distribution, but which

are nevertheless “effectively random.” The reason is that they are *computationally indistinguishable* from the uniform distribution.

Let us try to formalize this idea. Suppose we can randomly sample n -bit strings chosen according to a probability distribution P_n , and suppose that we want to know whether P_n is in fact the uniform distribution. One way to try to tell is to fix an efficiently computable function $f : \mathbb{I}_n \rightarrow \{0, 1\}$ and consider two experiments: one of the probability that $f(x) = 1$ when x is chosen with probability $P_n(x)$, and the other of the probability that $f(x) = 1$ when x is chosen with the uniform probability 2^{-n} . If there is a noticeable discrepancy between these two probabilities, then certainly P_n is not uniform. However, the converse is not true: it may be that P_n is far from uniform, but *no* efficiently computable function f can help us detect this. In that case, we say that P_n is *pseudorandom*.

This definition is both general and pragmatic. It refers to any efficient procedure that may be employed in an attempt to tell two distributions apart. And it is pragmatic because for any practical purpose a pseudorandom distribution is as good as a random one, for reasons we shall now explain.

Notice first that the behavior of any efficient probabilistic algorithm will be virtually unaffected if we replace its random source with a pseudorandom one. Why? Because if its behavior changed, then the algorithm itself would have efficiently distinguished between the random and pseudorandom sources, contradicting the definition of pseudorandomness!

Replacing uniform distributions by pseudorandom distributions is beneficial provided we can generate the latter using fewer resources. In this context, the resource we are trying hardest to save on is randomness. Suppose we have an efficiently computable function $\phi : \mathbb{I}_m \rightarrow \mathbb{I}_n$ and suppose that $n > m$. Then we can define a probability distribution on n -bit strings by choosing a random m -bit string x and computing $\phi(x)$. If this distribution is pseudorandom, then ϕ is called a *pseudorandom generator*. The random string x is called the *seed*, and if the generator stretches m -bit long seeds into strings of length $n = \ell(m)$, then we call the function ℓ the *stretch measure* of the generator. The larger the stretch measure, the better the generator is considered to be.

Of course, all this raises an important question: Do pseudorandom generators exist? It is to this question that we now turn.

14. Two probability distributions p_1 and p_2 are *statistically close* if they assign roughly the same probabilities: that is, if $p_1(E) \approx p_2(E)$ for every event E .

7.1.1 Hardness versus Randomness

There is an obvious connection between pseudorandom generators and computational difficulty, since the main property of a pseudorandom generator is that its output should be computationally hard to distinguish from a purely random string, even though the two distributions are significantly different. However, there is a much less obvious connection as well.

Theorem. *Pseudorandom generators exist if and only if one-way functions exist. Furthermore, if pseudorandom generators exist then they exist for any stretch measure that is a polynomial.*¹⁵

This theorem converts computational difficulty, or *hardness*, into pseudorandomness, and vice versa. Furthermore, its proof links computational indistinguishability to computational unpredictability, hinting that the computational difficulty is linked to randomness, or at least to the appearance of randomness.

The existence of pseudorandom generators has the remarkable consequence that probabilistic algorithms can be partially or even wholly *derandomized*. The basic idea is this. Suppose you have a probabilistic algorithm that computes a function f and requires n^c random bits (where n denotes the length of the input). Suppose that this algorithm outputs $f(x)$ with probability at least $\frac{2}{3}$. If you replace the random bits with n^c *pseudorandom* bits, generated from a seed of size m , then the behavior of the algorithm will hardly be affected. Therefore, if m is small, then you can do the same computation with only a small amount of randomness. If m is as small as $O(\log n)$, then it becomes feasible to check through *all* possible seeds. For close to two thirds of these, the algorithm outputs $f(x)$. But this means we can compute $f(x)$ deterministically and efficiently by taking a majority vote!

Can this actually be done? Can we use hardness to achieve the ultimate derandomization result, that $BPP = P$? The theory has developed to give essentially optimal answers to this question. Notice that if we wish to achieve an exponential stretch measure, we do not mind if the algorithm that performs the stretch takes exponential time (in the length of the seed). Such pseudorandom generators exist under very plausible hardness assumptions, such as the assumption that NP -complete problems require exponential-size

Boolean circuits. More generally, we have the following theorem.

Theorem. *If, for some constant $\epsilon > 0$, $S(SAT) > 2^{\epsilon n}$, then $BPP = P$. Moreover, SAT can be replaced by any problem computable in $2^{O(n)}$ -time.*

7.1.2 Pseudorandom Functions

Pseudorandom generators allow you to generate long pseudorandom sequences efficiently from short random seeds. Pseudorandom *functions* are even more powerful: if you are given a random seed of n bits, they provide you with an efficient way of computing a function $f : \mathbb{I}_n \rightarrow \{0, 1\}$ that is computationally indistinguishable from a random function. Thus, with just n bits of randomness, one has efficient access to 2^n bits that appear random. (Note that it is inefficient to scan through all these bits—what we are given is the ability to look at any one of them in polynomial time.)

It turns out that *pseudorandom functions can be constructed given any pseudorandom generator*, and that they have many applications (most notably in cryptography).

7.2 Cryptography

Cryptography has existed for millennia, but whereas in the past it was focused on one basic problem—that of providing secret communications—the modern computational theory of cryptography is interested in *all* tasks that involve several agents who each wish to obtain some information while preserving the secrecy of other information. An important priority besides *privacy* (that is, keeping secrets) is *resilience*: one would like guaranteed privacy even if one is not certain that the other participants are behaving honestly.

A good example to illustrate these difficulties is playing a game of poker over the telephone or e-mail. You are encouraged to ponder seriously how this might be done, and realize to what extent standard poker relies on human vision, physical implements like cards with opaque backs, etc., to protect privacy and prevent cheating.

In general, the goal of cryptography is to construct schemes, called *protocols*, that maintain *any* desired functionality (rules, privacy requirements, etc.), even in the face of malicious attempts to make these schemes deviate from this functionality. As with pseudorandomness, there are two key assumptions underlying the new theory. First, it is assumed that

15. In other words, if you can achieve a stretch measure $\ell(m) = m + 1$, then you can also achieve a stretch measure of $\ell(m) = m^c$ for any $c > 1$.

all parties, including the malicious adversaries, are computationally limited. Second, it is assumed that there are hard functions. Sometimes these are one-way functions, and sometimes they are yet stronger functions called “trapdoor permutations,” which also exist if integer factorization is hard.

This goal is an ambitious one, but it has been achieved. There is a result that says, roughly speaking, that *every functionality can be securely implemented*. This includes highly complex tasks such as playing poker over the phone, but also very basic ones such as secure communication, digital signatures (a digital analogue of handwritten signatures), collective coin flipping, auctions, elections, and the famous *millionaires’ problem*: how can two people interact to determine who is richer, without either of them learning anything further about the other’s wealth?

Let us very briefly hint at connections between cryptography and matters that we have already discussed. First of all, consider the very definition of the central notion of cryptography: that of a *secret*. If you have an n -bit string, then when should we say that it is completely secret? A natural definition would be that it is secret if nobody else has any information about it: that is, from anybody else’s point of view it is equally likely to be any of the 2^n -bit strings. However, in the new computational complexity theory, this is not the definition taken, since a *pseudorandom* n -bit string will, for all practical purposes, be just as secret.

The difference between the two definitions of a secret is huge. The point of cryptography is not just to have secrets (that is easy, just select a string at random) but actually to *use* them without giving away information. At first this seems impossible, since any nontrivial use of a secret n -bit string will cut down the set of possible strings that it might be, and therefore give away genuine information. However, if the new probability distribution over the possible strings (after the information has been given away) is pseudorandom, then this information *cannot feasibly be used*, since no efficient algorithm can tell the difference between a string that gives rise to the information you have revealed and a truly random string.

A famous example of this idea is given by the so-called *public-key encryption schemes*, such as RSA, which are described in detail in MATHEMATICS AND CRYPTOGRAPHY [VII.7] and in Goldreich (2004, chapter 5). In the RSA scheme, if a user, say Alice, wants to receive messages, she publishes a number N , called a *public key*, which is a product of two primes P and Q .

If you know N then you can encrypt any message, but to decrypt it you need to know P and Q . Thus, if integer factorization is hard, then only Alice can feasibly decrypt messages, even though P and Q are completely determined by N .

The generic problem about using secrets is one in which there are k parties, and each party has a string of bits. They are interested in the value of some efficiently computable function f that depends on all the strings of bits, but they would like to ascertain this without giving away any information about their own strings beyond what follows from the value of f . For example, in the case of the millionaires’ problem, there are two parties, each with a string that encodes their wealth. They would like a protocol that provides them with a single bit that tells them who is richer, but gives them no information beyond this. The precise formulation of this condition is an extension of the formulation of zero-knowledge proofs (presented in section 6.3.2). As hinted at earlier in this section, assuming the existence of trapdoor permutations, *every such multiparty computation can be performed without yielding anything beyond the designated outputs*.

Finally, we come to the issue of cheating. In the foregoing discussion, we did not worry about malicious behavior and focused on what participants may learn from the transcript of their interaction. But how can a player, Bob, say, be forced to act “as specified,” when his actions may depend partly on his secrets, which he does not want to reveal? The answer is closely related to zero-knowledge proofs. Essentially, each player whose turn it is to perform some computation is asked to prove to the others that he has acted as specified. This is a (mathematically boring) theorem and the standard proof is obvious (i.e., revealing all his secrets). But as we saw in our discussion of zero-knowledge proof systems in section 6.3.2, if a proof exists, then a zero-knowledge proof can be efficiently derived from it. Thus, *Bob can convince the others of his proper behavior without revealing anything about his secrets*.

8 The Tip of an Iceberg

Even within the topics reviewed above, many important notions and results have not been discussed, for space reasons. Furthermore, other important topics and even wide areas have not been mentioned at all.

The \mathcal{P} versus \mathcal{NP} question, as well as most of the discussion so far, focuses on a simplified view of the goals of (efficient) computations. Specifically, we have

PUP avoiding ‘each ... their’ would result in something so clumsy that Tim would prefer to keep things as they are. OK?

insisted on efficient procedures that *always* give the *exact* answer. However, in practice one may be content with less. For example, one may be happy with an efficient procedure that gives the correct answer for a large fraction of the instances. This will be useful if all instances are equally interesting, but that is typically not the case. On the other hand, demanding success under all input distributions gives back worst-case complexity. Between these two extremes is a useful and appealing theory of *average-case complexity* (see Goldreich 1997): one demands that algorithms succeed with high probability on every possible input distribution that can be *efficiently sampled*.

Another possible relaxation is settling for approximate answers. This can mean many things, and the best notion of approximation varies from context to context. For search problems, we may be satisfied with a solution that is close in some METRIC [III.58] to being valid (see Hochbaum (1996) and THE MATHEMATICS OF ALGORITHM DESIGN [VII.5]). For decision problems, we might ask how close the input is (again in some natural metric) to an instance in the set (see Ron 2001). And there is also approximate counting, which was discussed in section 6.2.

In this article we have focused on the *running time* of procedures. This is arguably the most important complexity measure, but it is not the only one. Another is the amount of *work space* consumed during the computation (see Sipser 1997). Another important issue is the extent to which a computation can be performed in parallel; that is, speeding up the computation by splitting the work among several computing devices, which are viewed as components of the same (parallel) machine and are provided with direct access to the same memory module. In addition to the parallel *time*, a fundamentally important complexity measure in such a case is the number of parallel computing devices used (see Karp and Ramachandran 1990).

Finally, there are several computational models that we have not discussed here. Models of *distributed computing* refer to distant computing devices, each given a local input, which may be viewed as a part of a global input. In typical studies one wishes to minimize the amount of communication between these devices (and certainly avoid the communication of the entire input). In addition to measures of communication complexity, a central issue is asynchrony (see Attiya and Welch 1998). The *communication complexity* of two-argument (and many-argument) functions is a measure of their “complexity” (see Kushilevitz and

Nisan 1996), but in these studies communication proportional to the length of the input is not ruled out (but rather appears frequently). While being “information theoretic” in nature, this model has many connections to complexity theory. Altogether different types of computational problems are investigated in the context of *computational learning theory* (see Kearns and Vazirani 1994) and of *online* algorithms (see Borodin and El-Yaniv 1998). Finally, QUANTUM COMPUTATION [III.76] investigates the possibility of using quantum mechanics to speed up computation (see Kitaev et al. 2002).

9 Concluding Remarks

We hope that this ultra-brief survey conveys the fascinating flavor of the concepts, results, and open problems that dominate the field of computational complexity. One important feature of the field we did not do justice to is the remarkable web of (often surprising) connections between different subareas, and its impact on progress.

For further details on sections 1–4 the reader is referred to standard textbooks such as Garey and Johnson (1979) and Sipser (1997). For further details on sections 5.1–5.3 the reader is referred to Boppana and Sipser (1990), Strassen (1990), and Beame and Pitassi (1998), respectively. For further details on sections 6 and 7 the reader is referred to Goldreich (1999) (and also to Goldreich (2001, 2004)).

Further Reading

- Attiya, H., and J. Welch. 1998. *Distributed Computing: Fundamentals, Simulations and Advanced Topics*. Columbus, OH: McGraw-Hill.
- Beame, P., and T. Pitassi. 1998. Propositional proof complexity: past, present, and future. *Bulletin of the European Association for Theoretical Computer Science* 65:66–89.
- Boppana, R., and M. Sipser. 1990. The complexity of finite functions. In *Handbook of Theoretical Computer Science*, volume A, *Algorithms and Complexity*, edited by J. van Leeuwen. Cambridge, MA: MIT Press/Elsevier.
- Borodin, A., and R. El-Yaniv. 1998. *On-line Computation and Competitive Analysis*. Cambridge: Cambridge University Press.
- Garey, M. R., and D. S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York: W. H. Freeman.
- Goldreich, O. 1997. Notes on Levin's theory of average-case complexity. *Electronic Colloquium on Computational Complexity*, TR97-058.

- . 1999. *Modern Cryptography, Probabilistic Proofs and Pseudorandomness*. Algorithms and Combinatorics Series, volume 17. New York: Springer.
- . 2001. *Foundation of Cryptography*, volume 1: *Basic Tools*. Cambridge: Cambridge University Press.
- . 2004. *Foundation of Cryptography*, volume 2: *Basic Applications*. Cambridge: Cambridge University Press.
- Hochbaum, D., ed. 1996. *Approximation Algorithms for NP-Hard Problems*. Boston, MA: PWS.
- Karp, R. M., and V. Ramachandran. 1990. Parallel algorithms for shared-memory machines. In *Handbook of Theoretical Computer Science*, volume A, *Algorithms and Complexity*, edited by J. van Leeuwen. Cambridge, MA: MIT Press/Elsevier.
- Kearns, M. J., and U. V. Vazirani. 1994. *An Introduction to Computational Learning Theory*. Cambridge, MA: MIT Press.
- Kitaev, A., A. Shen, and M. Vyalyi. 2002. *Classical and Quantum Computation*. Providence, RI: American Mathematical Society.
- Kushilevitz, E., and N. Nisan. 1996. *Communication Complexity*. Cambridge: Cambridge University Press.
- Ron, D. 2001. Property testing (a tutorial). In *Handbook on Randomized Computing*, volume II. Dordrecht: Kluwer.
- Shaltiel, R. 2002. Recent developments in explicit constructions of extractors. *Bulletin of the European Association for Theoretical Computer Science* 77:67–95.
- Sipser, M. 1997. *Introduction to the Theory of Computation*. Boston, MA: PWS.
- Strassen, V. 1990. Algebraic complexity theory. In *Handbook of Theoretical Computer Science*, volume A, *Algorithms and Complexity*, edited by J. van Leeuwen. Cambridge, MA: MIT Press/Elsevier.

IV.21 Numerical Analysis

Lloyd N. Trefethen

1 The Need for Numerical Computation

Everyone knows that when scientists and engineers need numerical answers to mathematical problems, they turn to computers. Nevertheless, there is a widespread misconception about this process.

The power of numbers has been extraordinary. It is often noted that the scientific revolution was set in motion when Galileo and others made it a principle that everything must be measured. Numerical measurements led to physical laws expressed mathematically, and, in the remarkable cycle whose fruits are all around us, finer measurements led to refined laws, which in turn led to better technology and still finer measurements. The day has long since passed when an advance

in the physical sciences could be achieved, or a significant engineering product developed, without numerical mathematics.

Computers certainly play a part in this story, yet there is a misunderstanding about what their role is. Many people imagine that scientists and mathematicians generate formulas, and then, by inserting numbers into these formulas, computers grind out the necessary results. The reality is nothing like this. What really goes on is a far more interesting process of execution of *algorithms*. In most cases the job could not be done even in principle by formulas, for most mathematical problems cannot be solved by a finite sequence of elementary operations. What happens instead is that fast algorithms quickly converge to “approximate” answers that are accurate to three or ten digits of precision, or a hundred. For a scientific or engineering application, such an answer may be as good as exact.

We can illustrate the complexities of exact versus approximate solutions by an elementary example. Suppose we have one polynomial of degree 4,

$$p(z) = c_0 + c_1z + c_2z^2 + c_3z^3 + c_4z^4,$$

and another of degree 5,

$$q(z) = d_0 + d_1z + d_2z^2 + d_3z^3 + d_4z^4 + d_5z^5.$$

It is well-known that there is an explicit formula that expresses the roots of p in terms of radicals (discovered by Ferrari around 1540), but no such formula for the roots of q (as shown by Ruffini and ABEL [VI.33] more than 250 years later; see THE INSOLUBILITY OF THE QUINTIC [V.24] for more details). Thus, in a certain philosophical sense the root-finding problems for p and q are utterly different. Yet in practice they hardly differ at all. If a scientist or a mathematician wants to know the roots of one of these polynomials, he or she will turn to a computer and get an answer to sixteen digits of precision in less than a millisecond. Did the computer use an explicit formula? In the case of q , the answer is certainly no, but what about p ? Maybe, maybe not. Most of the time, the user neither knows nor cares, and probably not one mathematician in a hundred could write down formulas for the roots of p from memory.

Here are three more examples of problems that can be solved in principle by a finite sequence of elementary operations, like rootfinding for p .

- (i) Linear equations: solve a system of n linear equations in n unknowns.

- (ii) Linear programming: minimize a linear function of n variables subject to m linear constraints.
- (iii) Traveling salesman problem: find the shortest tour between n cities.

And here are five that, like rootfinding for q , cannot generally be solved in this manner.

- (iv) Find an EIGENVALUE [I.3 §4.3] of an $n \times n$ matrix.
- (v) Minimize a function of several variables.
- (vi) Evaluate an integral.
- (vii) Solve an ordinary differential equation (ODE).
- (viii) Solve a partial differential equation (PDE).

Can we conclude that (i)–(iii) will be easier than (iv)–(viii) in practice? Absolutely not. Problem (iii) is usually very hard indeed if n is, say, in the hundreds or thousands. Problems (vi) and (vii) are usually rather easy, at least if the integral is in one dimension. Problems (i) and (iv) are of almost exactly the same difficulty: easy when n is small, like 100, and often very hard when n is large, like 1 000 000. In fact, in these matters philosophy is such a poor guide to practice that, for each of the three problems (i)–(iii), when n and m are large one often ignores the exact solution and uses approximate (but fast!) methods instead.

Numerical analysis is the study of algorithms for solving the problems of continuous mathematics, by which we mean problems involving real or complex variables. (This definition includes problems like linear programming and the traveling salesman problem posed over the real numbers, but not their discrete analogues.) In the remainder of this article we shall review some of its main branches, past accomplishments, and possible future trends.

2 A Brief History

Throughout history, leading mathematicians have been involved with scientific applications, and in many cases this has led to the discovery of numerical algorithms still in use today. GAUSS [VI.26], as usual, is an outstanding example. Among many other contributions, he made crucial advances in least-squares data fitting (1795), systems of linear equations (1809), and numerical quadrature (1814), as well as inventing THE FAST FOURIER TRANSFORM [III.26] (1805), though the last did not become widely known until its rediscovery by Cooley and Tukey in 1965.

Around 1900, the numerical side of mathematics started to become less conspicuous in the activities

of research mathematicians. This was a consequence of the growth of mathematics generally and of great advances in fields in which, for technical reasons, mathematical rigor had to be the heart of the matter. For example, many advances of the early twentieth century sprang from mathematicians' new ability to reason rigorously about infinity, a subject relatively far from numerical calculation.

A generation passed, and in the 1940s the computer was invented. From this moment numerical mathematics began to explode, but now mainly in the hands of specialists. New journals were founded such as *Mathematics of Computation* (1943) and *Numerische Mathematik* (1959). The revolution was sparked by hardware, but it included mathematical and algorithmic developments that had nothing to do with hardware. In the half-century from the 1950s, machines sped up by a factor of around 10^9 , but so did the best algorithms known for some problems, generating a combined increase in speed of almost incomprehensible scale.

Half a century on, numerical analysis has grown into one of the largest branches of mathematics, the specialty of thousands of researchers who publish in dozens of mathematical journals as well as applications journals across the sciences and engineering. Thanks to the efforts of these people going back many decades, and thanks to ever more powerful computers, we have reached a point where most of the classical mathematical problems of the physical sciences can be solved numerically to high accuracy. Most of the algorithms that make this possible were invented since 1950.

Numerical analysis is built on a strong foundation: the mathematical subject of *approximation theory*. This field encompasses classical questions of interpolation, series expansions, and HARMONIC ANALYSIS [IV.11] associated with NEWTON [VI.14], FOURIER [VI.25], Gauss, and others; semiclassical problems of polynomial and rational minimax approximation associated with names such as CHEBYSHEV [VI.45] and Bernstein; and major newer topics, including splines, radial basis functions, and WAVELETS [VII.3]. We shall not have space to address these subjects, but in almost every area of numerical analysis it is a fact that, sooner or later, the discussion comes down to approximation theory.

3 Machine Arithmetic and Rounding Errors

It is well-known that computers cannot represent real or complex numbers exactly. A quotient like $1/7$ eval-

uated on a computer, for example, will normally yield an inexact result. (It would be different if we designed machines to work in base 7!) Computers approximate real numbers by a system of *floating-point arithmetic*, in which each number is represented in a digital equivalent of scientific notation, so that the scale does not matter unless the number is so huge or tiny as to cause overflow or underflow. Floating-point arithmetic was invented by Konrad Zuse in Berlin in the 1930s, and by the end of the 1950s it was standard across the computer industry.

Until the 1980s, different computers had widely different arithmetic properties. Then, in 1985, after years of discussion, the IEEE (Institute of Electrical and Electronics Engineers) standard for binary floating-point arithmetic was adopted, or *IEEE arithmetic* for short. This standard has subsequently become nearly universal on processors of many kinds. An IEEE (double precision) real number consists of a 64-bit word divided into 53 bits for a signed fraction in base 2 and 11 bits for a signed exponent. Since $2^{-53} \approx 1.1 \times 10^{-16}$, IEEE numbers represent the numbers of the real line to a relative accuracy of about 16 digits. Since $2^{\pm 2^{10}} \approx 10^{\pm 308}$, this system works for numbers up to about 10^{308} and down to about 10^{-308} .

Computers do not merely represent numbers, of course; they perform operations on them such as addition, subtraction, multiplication, and division, and more complicated results are obtained from sequences of these elementary operations. In floating-point arithmetic, the computed result of each elementary operation is almost exactly correct in the following sense: if “ $*$ ” is one of these four operations in its ideal form and “ \odot ” is the same operation as realized on the computer, then for any floating-point numbers x and y , assuming that there is no underflow or overflow,

$$x \odot y = (x * y)(1 + \varepsilon).$$

Here ε is a very small quantity, no greater in absolute value than a number known as *machine epsilon*, denoted by $\varepsilon_{\text{mach}}$, that measures the accuracy of the computer. In the IEEE system, $\varepsilon_{\text{mach}} = 2^{-53} \approx 1.1 \times 10^{-16}$.

Thus, on a computer, the interval $[1, 2]$, for example, is approximated by about 10^{16} numbers. It is interesting to compare the fineness of this discretization with that of the discretizations of physics. In a handful of solid or liquid or a balloonful of gas, the number of atoms or molecules in a line from one point to another

is on the order of 10^8 (the cube root of Avogadro’s number). Such a system behaves enough like a continuum to justify our definitions of physical quantities such as density, pressure, stress, strain, and temperature. Computer arithmetic, however, is more than a million times finer than this. Another comparison with physics concerns the precision to which fundamental constants are known, such as (roughly) 4 digits for the gravitational constant G , 7 digits for Planck’s constant h and the elementary charge e , and 12 digits for the ratio μ_e/μ_B of the magnetic moment of the electron to the Bohr magneton. At present, almost nothing in physics is known to more than 12 or 13 digits of accuracy. Thus IEEE numbers are orders of magnitude more precise than any number in science. (Of course, purely mathematical quantities like π are another matter.)

In two senses, then, floating-point arithmetic is far closer to its ideal than is physics. It is a curious phenomenon that, nevertheless, it is floating-point arithmetic rather than the laws of physics that is widely regarded as an ugly and dangerous compromise. Numerical analysts themselves are partly to blame for this perception. In the 1950s and 1960s, the founding fathers of the field discovered that inexact arithmetic can be a source of danger, causing errors in results that “ought” to be right. The source of such problems is *numerical instability*: that is, the amplification of rounding errors from microscopic to macroscopic scale by certain modes of computation. These men, including VON NEUMANN [VI.91], Wilkinson, Forsythe, and Henrici, took great pains to publicize the risks of careless reliance on machine arithmetic. These risks are very real, but the message was communicated all too successfully, leading to the current widespread impression that the main business of numerical analysis is coping with rounding errors. In fact, the main business of numerical analysis is designing algorithms that converge quickly; rounding-error analysis, while often a part of the discussion, is rarely the central issue. If rounding errors vanished, 90% of numerical analysis would remain.

PUP: Tim prefers ‘is’. OK?

4 Numerical Linear Algebra

Linear algebra became a standard topic in undergraduate mathematics curriculums in the 1950s and 1960s, and has remained there ever since. There are several reasons for this, but I think one is at the bottom of it: the importance of linear algebra has exploded since the arrival of computers.

T&T note: must check that this star-in-a-circle still looks OK before CRC.

The starting point of this subject is *Gaussian elimination*, a procedure that can solve n linear equations in n unknowns using on the order of n^3 arithmetic operations. Equivalently, it solves equations of the form $Ax = b$, where A is an $n \times n$ matrix and x and b are column vectors of size n . Gaussian elimination is invoked on computers around the world almost every time a system of linear equations is solved. Even if n is as large as 1000, the time required is well under a second on a typical 2008 desktop machine. The idea of elimination was first discovered by Chinese scholars about 2000 years ago, and more recent contributors include LAGRANGE [VI.22], Gauss, and JACOBI [VI.35]. The modern way of describing such algorithms, however, was apparently introduced as late as the 1930s. Suppose that, say, α times the first row of A is subtracted from the second row. This operation can be interpreted as the multiplication of A on the left by the lower-triangular matrix M_1 consisting of the identity with the additional nonzero entry $m_{21} = -\alpha$. Further analogous row operations correspond to further multiplications on the left by lower-triangular matrices M_j . If k steps convert A to an upper-triangular matrix U , then we have $MA = U$ with $M = M_k \cdots M_2 M_1$, or, upon setting $L = M^{-1}$,

$$A = LU.$$

Here L is unit lower-triangular, that is, lower-triangular with all its diagonal entries equal to 1. Since U represents the target structure and L encodes the operations carried out to get there, we can say that Gaussian elimination is a process of *lower-triangular upper-triangularization*.

Many other algorithms of numerical linear algebra are also based on writing a matrix as a product of matrices that have special properties. To borrow a phrase from biology, we may say that this field has a central dogma:

$$\text{algorithms} \longleftrightarrow \text{matrix factorizations.}$$

In this framework we can quickly describe the next algorithm that needs to be considered. Not every matrix has an LU factorization; a 2×2 counterexample is the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Soon after computers came into use it was observed that even for matrices that do have LU factorizations, the pure form of Gaussian elimination is unstable, amplifying rounding errors by potentially large amounts. Stability can be achieved by interchanging

rows during the elimination in order to bring maximal entries to the diagonal, a process known as *pivoting*. Since pivoting acts on rows, it again corresponds to a multiplication of A by other matrices on the left. The matrix factorization corresponding to Gaussian elimination with pivoting is

$$PA = LU,$$

where U is upper-triangular, L is unit lower-triangular, and P is a permutation matrix, i.e., the identity matrix with permuted rows. If the permutations are chosen to bring the largest entry below the diagonal in column k to the (k, k) position before the k th elimination step, then L has the additional property $|\ell_{ij}| \leq 1$ for all i and j .

The discovery of pivoting came quickly, but its theoretical analysis has proved astonishingly hard. In practice, pivoting makes Gaussian elimination almost perfectly stable, and it is routinely done by almost all computer programs that need to solve linear systems of equations. Yet it was realized in around 1960 by Wilkinson and others that for certain exceptional matrices, Gaussian elimination is still unstable, even with pivoting. The lack of an explanation of this discrepancy represents an embarrassing gap at the heart of numerical analysis. Experiments suggest that the fraction of matrices (for example, among random matrices with independent normally distributed entries) for which Gaussian elimination amplifies rounding errors by a factor greater than $\rho n^{1/2}$ is in a certain sense exponentially small as a function of ρ as $\rho \rightarrow \infty$, where n is the dimension, but a theorem to this effect has never been proved.

Meanwhile, beginning in the late 1950s, the field of numerical linear algebra expanded in another direction: the use of algorithms based on ORTHOGONAL [III.52 §3] or UNITARY [III.52 §3] matrices, that is, real matrices with $Q^{-1} = Q^T$ or complex ones with $Q^{-1} = Q^*$, where Q^* denotes the conjugate transpose. The starting point of such developments is the idea of *QR factorization*. If A is an $m \times n$ matrix with $m \geq n$, a QR factorization of A is a product

$$A = QR,$$

where Q has orthonormal columns and R is upper-triangular. One can interpret this formula as a matrix expression of the familiar idea of *Gram-Schmidt orthogonalization*, in which the columns q_1, q_2, \dots of Q are determined one after another. These column operations correspond to multiplication of A on the right by elementary upper-triangular matrices. One could say

that the Gram-Schmidt algorithm aims for Q and gets R as a by-product, and is thus a process of *triangular orthogonalization*. A big event was when Householder showed in 1958 that a dual strategy of *orthogonal triangularization* is more effective for many purposes. In this approach, by applying a succession of elementary matrix operations each of which reflects \mathbb{R}^m across a hyperplane, one reduces A to upper-triangular form via orthogonal operations: one aims at R and gets Q as a by-product. The Householder method turns out to be more stable numerically, because orthogonal operations preserve norms and thus do not amplify the rounding errors introduced at each step.

From the QR factorization sprang a rich collection of linear algebra algorithms in the 1960s. The QR factorization can be used by itself to solve least-squares problems and construct orthonormal bases. More remarkable is its use as a step in other algorithms. In particular, one of the central problems of numerical linear algebra is the determination of the eigenvalues and eigenvectors of a square matrix A . If A has a complete set of eigenvectors, then by forming a matrix X whose columns are these eigenvectors and a diagonal matrix D whose diagonal entries are the corresponding eigenvalues, we obtain

$$AX = XD,$$

and hence, since X is nonsingular,

$$A = XDX^{-1},$$

the *eigenvalue decomposition*. In the special case in which A is HERMITIAN [III.52 §3], a complete set of orthonormal eigenvectors always exists, giving

$$A = QDQ^*,$$

where Q is unitary. The standard algorithm for computing these factorizations was developed in the early 1960s by Francis, Kublanovskaya, and Wilkinson: the *QR algorithm*. Because polynomials of degree 5 or more cannot be solved by a formula, we know that eigenvalues cannot generally be computed in closed form. The QR algorithm is therefore necessarily an iterative one, involving a sequence of QR factorizations that is in principle infinite. Nevertheless, its convergence is extraordinarily rapid. In the symmetric case, for a typical matrix A , the QR algorithm converges *cubically*, in the sense that at each step the number of correct digits in one of the eigenvalue-eigenvector pairs approximately triples.

The QR algorithm is one of the great triumphs of numerical analysis, and its impact through widely used

software products has been enormous. Algorithms and analysis based on it led in the 1960s to computer codes in Algol and Fortran and later to the software library EISPACK ("Eigensystem Package") and its descendant LAPACK. The same methods have also been incorporated in general-purpose numerical libraries such as the NAG, IMSL, and *Numerical Recipes* collections, and in problem-solving environments such as MATLAB, Maple, and Mathematica. These developments have been so successful that the computation of matrix eigenvalues long ago became a "black box" operation for virtually every scientist, with nobody but a few specialists knowing the details of how it is done. A curious related story is that EISPACK's relative LINPACK for solving linear systems of equations took on an unexpected function: it became the original basis for the benchmarks that all computer manufacturers run to test the speed of their computers. If a supercomputer is lucky enough to make the TOP500 list, updated twice a year since 1993, it is because of its prowess in solving certain matrix problems $Ax = b$ of dimensions ranging from 100 into the millions.

The eigenvalue decomposition is familiar to all mathematicians, but the development of numerical linear algebra has also brought its younger cousin onto the scene: the *singular value decomposition* (SVD). The SVD was discovered by Beltrami, JORDAN [VI.52], and SYLVESTER [VI.42] in the late nineteenth century, and made famous by Golub and other numerical analysts beginning in around 1965. If A is an $m \times n$ matrix with $m \geq n$, an SVD of A is a factorization

$$A = U\Sigma V^*,$$

where U is $m \times n$ with orthonormal columns, V is $n \times n$ and unitary, and Σ is diagonal with diagonal entries $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$. One could compute the SVD by relating it to the eigenvalue problems for AA^* and A^*A , but this proves numerically unstable; a better approach is to use a variant of the QR algorithm that does not square A . Computing the SVD is the standard route to determining the NORM [III.64] $\|A\| = \sigma_1$ (here $\|\cdot\|$ is the HILBERT SPACE [III.37] or "2" norm), the norm of the inverse $\|A^{-1}\| = 1/\sigma_n$ in the case where A is square and nonsingular, or their product, known as the *condition number*,

$$\kappa(A) = \|A\| \|A^{-1}\| = \sigma_1/\sigma_n.$$

It is also a step in an extraordinary variety of further computational problems including rank-deficient least-squares, computation of ranges and nullspaces,

determination of ranks, “total least-squares,” low-rank approximation, and determination of angles between subspaces.

All the discussion above concerns “classical” numerical linear algebra, born in the period 1950–75. The ensuing quarter-century brought in a whole new set of tools: methods for large-scale problems based on *Krylov subspace iterations*. The idea of these iterations is as follows. Suppose a linear algebra problem is given that involves a matrix of large dimension, say $n \gg 1000$. The solution may be characterized as the vector $x \in \mathbb{R}^n$ that satisfies a certain variational property such as minimizing $\frac{1}{2}x^T Ax - x^T b$ (for solving $Ax = b$ if A is symmetric positive definite) or being a stationary point of $(x^T Ax)/(x^T x)$ (for solving $Ax = \lambda x$ if A is symmetric). Now if K_k is a k -dimensional subspace of \mathbb{R}^n with $k \ll n$, then it may be possible to solve the same variational problem much more quickly in that subspace. The magical choice of K_k is a *Krylov subspace*

$$K_k(A, q) = \text{span}(q, Aq, \dots, A^{k-1}q)$$

for an initial vector q . For reasons that have fascinating connections with approximation theory, solutions in these subspaces often converge very rapidly to the exact solution in \mathbb{R}^n as k increases, if the eigenvalues of A are favorably distributed. For example, it is often possible to solve a matrix problem involving 10^5 unknowns to ten-digit precision in just a few hundred iterations. The speedup compared with the classical algorithms may be a factor of thousands.

Krylov subspace iterations originated with the conjugate gradient and Lanczos iterations published in 1952, but in those years computers were not powerful enough to solve problems of a large enough scale for the methods to be competitive. They took off in the 1970s with the work of Reid and Paige and especially van der Vorst and Meijerink, who made famous the idea of *preconditioning*. In preconditioning a system $Ax = b$, one replaces it by a mathematically equivalent system such as

$$MAx = Mb$$

for some nonsingular matrix M . If M is well chosen, the new problem involving MA may have favorably distributed eigenvalues and a Krylov subspace iteration may solve it quickly.

Since the 1970s, preconditioned matrix iterations have emerged as an indispensable tool of computational science. As one indication of their prominence we may note that in 2001, Thomson ISI announced that the most heavily cited article in all of mathematics in the

1990s was the 1989 paper by van der Vorst introducing Bi-CGStab, a generalization of conjugate gradients for nonsymmetric matrices.

Finally, we must mention the biggest unsolved problem in numerical analysis. Can an arbitrary $n \times n$ matrix A be inverted in $O(n^\alpha)$ operations for every $\alpha > 2$? (The problems of solving a system $Ax = b$ or computing a matrix product AB are equivalent.) Gaussian elimination has $\alpha = 3$, and the exponent shrinks as far as 2.376 for certain recursive (though impractical) algorithms published by Coppersmith and Winograd in 1990. Is there a “fast matrix inverse” in store for us?

5 Numerical Solution of Differential Equations

Long before much attention was paid to linear algebra, mathematicians developed numerical methods to solve problems of analysis. The problem of numerical integration or *quadrature* goes back to Gauss and NEWTON [VI.14], and even to ARCHIMEDES [VI.3]. The classic quadrature formulas are derived from the idea of interpolating data at $n + 1$ points by a polynomial of degree n , then integrating the polynomial exactly. Equally spaced interpolation points give the *Newton-Cotes formulas*, which are useful for small degrees but diverge at a rate as high as 2^n as $n \rightarrow \infty$: the *Runge phenomenon*. If the points are chosen optimally, then the result is *Gauss quadrature*, which converges rapidly and is numerically stable. It turns out that these optimal points are roots of Legendre polynomials, which are clustered near the endpoints. (A proof is sketched in SPECIAL FUNCTIONS [III.87].) Equally good for most purposes is *Clenshaw-Curtis quadrature*, where the interpolation points become $\cos(j\pi/n)$, $0 \leq j \leq n$. This quadrature method is also stable and rapidly convergent, and unlike Gauss quadrature can be executed in $O(n \log n)$ operations by the fast Fourier transform. The explanation of why clustered points are necessary for effective quadrature rules is related to the subject of potential theory.

Around 1850 another problem of analysis began to get attention: the solution of ODEs. The *Adams formulas* are based on polynomial interpolation in equally spaced points, which in practice typically number fewer than ten. These were the first of what are now called *multistep methods* for the numerical solution of ODEs. The idea here is that for an initial value problem $u' = f(t, u)$ with independent variable $t > 0$, we pick a small time step $\Delta t > 0$ and consider a finite set of time values

$$t_n = n\Delta t, \quad n \geq 0.$$

We then replace the ODE by an algebraic approximation that enables us to calculate a succession of approximate values

$$v^n \approx u(t_n), \quad n \geq 0.$$

(The superscript here is just a superscript, not a power.) The simplest such approximate formula, going back to EULER [VI.19], is

$$v^{n+1} = v^n + \Delta t f(t_n, v^n),$$

or, using the abbreviation $f^n = f(t_n, v^n)$,

$$v^{n+1} = v^n + \Delta t f^n.$$

Both the ODE itself and its numerical approximation may involve one equation or many, in which case $u(t, x)$ and v^n become vectors of an appropriate dimension. The Adams formulas are higher-order generalizations of Euler's formula that are much more efficient at generating accurate solutions. For example, the fourth-order Adams-Bashforth formula is

$$v^{n+1} = v^n + \frac{1}{24} \Delta t (55f^n - 59f^{n-1} + 37f^{n-2} - 9f^{n-3}).$$

The term "fourth-order" reflects a new element in the numerical treatment of problems of analysis: the appearance of questions of convergence as $\Delta t \rightarrow 0$. The formula above is of fourth order in the sense that it will normally converge at the rate $O((\Delta t)^4)$. The orders employed in practice are most often in the range 3–6, enabling excellent accuracy for all kinds of computations, typically in the range of 3–10 digits, and higher-order formulas are occasionally used when still more accuracy is needed.

Most unfortunately, the habit in the numerical analysis literature is to speak not of the *convergence* of these magnificently efficient methods, but of their *error*, or more precisely their *discretization* or *truncation error* as distinct from rounding error. This ubiquitous language of error analysis is dismal in tone, but seems ineradicable.

At the turn of the twentieth century, the second great class of ODE algorithms, known as *Runge-Kutta* or *one-step methods*, was developed by Runge, Heun, and Kutta. For example, here are the formulas of the famous fourth-order Runge-Kutta method, which advance a numerical solution (again scalar or system) from time step t_n to t_{n+1} with the aid of four evaluations of the

function f :

$$\begin{aligned} a &= \Delta t f(t_n, v^n), \\ b &= \Delta t f(t_n + \frac{1}{2}\Delta t, v^n + \frac{1}{2}a), \\ c &= \Delta t f(t_n + \frac{1}{2}\Delta t, v^n + \frac{1}{2}b), \\ d &= \Delta t f(t_n + \Delta t, v^n + c), \\ v^{n+1} &= v^n + \frac{1}{6}(a + 2b + 2c + d). \end{aligned}$$

Runge-Kutta methods tend to be easier to implement but sometimes harder to analyze than multistep formulas. For example, for any s , it is a trivial matter to derive the coefficients of the s -step Adams-Bashforth formula, which has order of accuracy $p = s$. For Runge-Kutta methods, by contrast, there is no simple relationship between the number of "stages" (i.e., function evaluations per step) and the attainable order of accuracy. The classical methods with $s = 1, 2, 3, 4$ were known to Kutta in 1901 and have order $p = s$, but it was not until 1963 that it was proved that $s = 6$ stages are required to achieve order $p = 5$. The analysis of such problems involves beautiful mathematics from graph theory and other areas, and a key figure in this area since the 1960s has been John Butcher. For orders $p = 6, 7, 8$ the minimal numbers of stages are $s = 7, 9, 11$, while for $p > 8$ exact minima are not known. Fortunately, these higher orders are rarely needed for practical purposes.

When computers began to be used to solve differential equations after World War II, a phenomenon of the greatest practical importance appeared: once again, *numerical instability*. As before, this phrase refers to the unbounded amplification of local errors by a computational process, but now the dominant local errors are usually those of discretization rather than rounding. Instability typically manifests itself as an oscillatory error in the computed solution that blows up exponentially as more numerical steps are taken. One mathematician concerned with this effect was Germund Dahlquist. Dahlquist saw that the phenomenon could be analyzed with great power and generality, and some people regard the appearance of his 1956 paper as one of the events marking the birth of modern numerical analysis. This landmark paper introduced what might be called the *fundamental theorem of numerical analysis*:

$$\text{consistency} + \text{stability} = \text{convergence}.$$

The theory is based on precise definitions of these three notions along the following lines. *Consistency* is the property that the discrete formula has locally positive

order of accuracy and thus models the right ODE. *Stability* is the property that errors introduced at one time step cannot grow unboundedly at later times. *Convergence* is the property that as $\Delta t \rightarrow 0$, in the absence of rounding errors, the numerical solution converges to the correct result. Before Dahlquist's paper, the idea of an equivalence of stability and convergence was perhaps in the air in the sense that practitioners realized that if a numerical scheme was not unstable, then it would probably give a good approximation to the right answer. His theory gave rigorous form to that idea for a wide class of numerical methods.

As computer methods for ODEs were being developed, the same was happening for the much bigger subject of PDEs. Discrete numerical methods for solving PDEs had been invented around 1910 by Richardson for applications in stress analysis and meteorology, and further developed by Southwell; in 1928 there was also a theoretical paper on finite-difference methods by COURANT [VI.83], Friedrichs, and Lewy. But although the Courant–Friedrichs–Lewy work later became famous, the impact of these ideas before computers came along was limited. After that point the subject developed quickly. Particularly influential in the early years was the group of researchers around von Neumann at the Los Alamos laboratory, including the young Peter Lax.

Just as for ODEs, von Neumann and his colleagues discovered that some numerical methods for PDEs were subject to catastrophic instabilities. For example, to solve the linear wave equation $u_t = u_x$ numerically we may pick space and time steps Δx and Δt for a regular grid,

$$x_j = j\Delta x, \quad t_n = n\Delta t, \quad j, n \geq 0,$$

and replace the PDE by algebraic formulas that compute a succession of approximate values:

$$v_j^n \approx u(t_n, x_j), \quad j, n \geq 0.$$

A well-known discretization for this purpose is the *Lax–Wendroff formula*:

$$v_j^{n+1} = v_j^n + \frac{1}{2}\lambda(v_{j+1}^n - v_{j-1}^n) + \frac{1}{2}\lambda^2(v_{j+1}^n - 2v_j^n + v_{j-1}^n),$$

where $\lambda = \Delta t/\Delta x$, which can be generalized to nonlinear systems of hyperbolic conservation laws in one dimension. For $u_t = u_x$, if λ is held fixed at a value less than or equal to 1, the method will converge to the correct solution as $\Delta x, \Delta t \rightarrow 0$ (ignoring rounding errors). If λ is greater than 1, on the other hand, it will explode. Von Neumann and others realized that the presence or absence of such instabilities could be tested, at least

for linear constant-coefficient problems, by discrete FOURIER ANALYSIS [III.27] in x : “von Neumann analysis.” Experience indicated that, as a practical matter, a method would succeed if it was not unstable. A theory soon appeared that gave rigor to this observation: the *Lax equivalence theorem*, published by Lax and Richtmyer in 1956, the same year as Dahlquist's paper. Many details were different—this theory was restricted to linear equations whereas Dahlquist's theory for ODEs also applied to nonlinear ones—but broadly speaking the new result followed the same pattern of equating convergence to consistency plus stability. Mathematically, the key point was the uniform boundedness principle.

In the half-century since von Neumann died, the Lax–Wendroff formula and its relatives have grown into a breathtakingly powerful subject known as *computational fluid dynamics*. Early treatments of linear and nonlinear equations in one space dimension soon moved to two dimensions and eventually to three. It is now a routine matter to solve problems involving millions of variables on computational grids with hundreds of points in each of three directions. The equations are linear or nonlinear; the grids are uniform or nonuniform, often adaptively refined to give special attention to boundary layers and other fast-changing features; the applications are everywhere. Numerical methods were used first to model airfoils, then whole wings, then whole aircraft. Engineers still use wind tunnels, but they rely more on computations.

Many of these successes have been facilitated by another numerical technology for solving PDEs that emerged in the 1960s from diverse roots in engineering and mathematics: finite elements. Instead of approximating a differential operator by a difference quotient, finite-element methods approximate the solution itself by functions f that can be broken up into simple pieces. For instance, one might partition the domain of f into elementary sets such as triangles or tetrahedra and insist that the restriction of f to each piece is a polynomial of small degree. The solution is obtained by solving a variational form of the PDE within the corresponding finite-dimensional subspace, and there is often a guarantee that the computed solution is optimal within that subspace. Finite-element methods have taken advantage of tools of functional analysis to develop to a very mature state. These methods are known for their flexibility in handling complicated geometries, and in particular they are entirely dominant in applications in structural mechanics and civil engineering. The number of books and articles that

have been published about finite-element methods is in excess of 10 000.

In the vast and mature field of numerical solution of PDEs, what aspect of the current state of the art would most surprise Richardson or Courant, Friedrichs, and Lewy? I think it is the universal dependence on exotic algorithms of linear algebra. The solution of a large-scale PDE problem in three dimensions may require a system of a million equations to be solved at each time step. This may be achieved by a GMRES matrix iteration that utilizes a finite-difference preconditioner implemented by a Bi-CGSTab iteration relying on another multigrid preconditioner. Such stacking of tools was surely not imagined by the early computer pioneers. The need for it ultimately traces to numerical instability, for as Crank and Nicolson first noted in 1947, the crucial tool for combating instability is the use of *implicit formulas*, which couple together unknowns at the new time step t_{n+1} , and it is in implementing this coupling that solutions of systems of equations are required.

Here are some examples that illustrate the successful reliance of today's science and engineering on the numerical solution of PDEs: chemistry (SCHRÖDINGER EQUATION [III.85]); structural mechanics (equations of elasticity); weather prediction (geostrophic equations); turbine design (NAVIER-STOKES EQUATIONS [III.23]); acoustics (Helmholtz equation); telecommunications (MAXWELL'S EQUATIONS [IV.13 §1.1]); cosmology (Einstein equations); oil discovery (migration equations); groundwater remediation (Darcy's law); integrated circuit design (drift diffusion equations); tsunami modeling (shallow-water equations); optical fibers (NON-LINEAR WAVE EQUATIONS [III.51]); image enhancement (Perona-Malik equation); metallurgy (Cahn-Hilliard equation); pricing financial options (BLACK-SCHOLES EQUATION [VII.9 §2]).

6 Numerical Optimization

The third great branch of numerical analysis is optimization, that is, the minimization of functions of several variables and the closely related problem of solution of nonlinear systems of equations. The development of optimization has been somewhat independent of that of the rest of numerical analysis, carried forward in part by a community of scholars with close links to operations research and economics.

Calculus students learn that a smooth function may achieve an extremum at a point of zero derivative, or at

a boundary. The same two possibilities characterize the two big strands of the field of optimization. At one end there are problems of finding interior zeros and minima of unconstrained nonlinear functions by methods related to multivariate calculus. At the other are problems of linear programming, where the function to be minimized is linear and therefore easy to understand, and all the challenge is in the boundary constraints.

Unconstrained nonlinear optimization is an old subject. Newton introduced the idea of approximating functions by the first few terms of what we now call their Taylor series; indeed, Arnol'd has argued that Taylor series were Newton's "main mathematical discovery." To find a zero x_* of a function F of a real variable x , everyone knows the idea of *Newton's method*: at the k th step, given an estimate $x^{(k)} \approx x_*$, use the derivative $F'(x^{(k)})$ to define a linear approximation from which to derive a better estimate $x^{(k+1)}$:

$$x^{(k+1)} = x^{(k)} - F(x^{(k)})/F'(x^{(k)}).$$

Newton (1669) and Raphson (1690) applied this idea to polynomials, and Simpson (1740) generalized it to other functions F and to systems of two equations. In today's language, for a system of n equations in n unknowns, we regard F as an n -vector whose derivative at a point $x^{(k)} \in \mathbb{R}^n$ is the $n \times n$ Jacobian matrix with entries

$$J_{ij}(x^{(k)}) = \frac{\partial F_i}{\partial x_j}(x^{(k)}), \quad 1 \leq i, j \leq n.$$

This matrix defines a linear approximation to $F(x)$ that is accurate for $x \approx x^{(k)}$. Newton's method then takes the matrix form

$$x^{(k+1)} = x^{(k)} - (J(x^{(k)}))^{-1}F(x^{(k)}),$$

which in practice means that to get $x^{(k+1)}$ from $x^{(k)}$, we solve a linear system of equations:

$$J(x^{(k)})(x^{(k+1)} - x^{(k)}) = -F(x^{(k)}).$$

As long as J is Lipschitz continuous and nonsingular at x_* and the initial guess is good enough, the convergence of this iteration is quadratic:

$$\|x^{(k+1)} - x_*\| = O(\|x^{(k)} - x_*\|^2). \quad (1)$$

Students often think it might be a good idea to develop formulas to enhance the exponent in this estimate to 3 or 4. However, this is an illusion. Taking two steps at a time of a quadratically convergent algorithm yields a quartically convergent one, so the difference in efficiency between quadratic and quartic is at best a constant factor. The same goes if the exponent 2, 3, or 4 is replaced by any other number greater than 1.

PUP: this is correctly set.

The true distinction is between all of these algorithms that converge *superlinearly*, of which Newton's method is the prototype, and those that converge *linearly* or *geometrically*, where the exponent is just 1.

From the point of view of multivariate calculus, it is a small step from solving a system of equations to minimizing a scalar function f of a variable $x \in \mathbb{R}^n$: to find a (local) minimum, we seek a zero of the gradient $g(x) = \nabla f(x)$, an n -vector. The derivative of g is the Jacobian matrix known as the *Hessian* of f , with entries

$$H_{ij}(x^{(k)}) = \frac{\partial^2 f}{\partial x_i \partial x_j}(x^{(k)}), \quad 1 \leq i, j \leq n,$$

and one may utilize it just as before in a Newton iteration to find a zero of $g(x)$, the new feature being that a Hessian is always symmetric.

Though the Newton formulas for minimization and finding zeros were already established, the arrival of computers created a new field of numerical optimization. One of the obstacles quickly encountered was that Newton's method often fails if the initial guess is not good. This problem has been comprehensively addressed both practically and theoretically by the algorithmic technologies known as *line searches* and *trust regions*.

For problems with more than a few variables, it also quickly became clear that the cost of evaluating Jacobians or Hessians at every step could be exorbitant. Faster methods were needed that might make use of inexact Jacobians or Hessians and/or inexact solutions of the associated linear equations, while still achieving superlinear convergence. An early breakthrough of this kind was the discovery of *quasi-Newton methods* in the 1960s by Broyden, Davidon, Fletcher, and Powell, in which partial information is used to generate steadily improving estimates of the true Jacobian or Hessian or its matrix factors. An illustration of the urgency of this subject at the time is the fact that in 1970 the optimal rank-two symmetric positive-definite quasi-Newton updating formula was published independently by no fewer than four different authors, namely Broyden, Fletcher, Goldfarb, and Shanno; their discovery has been known ever since as the *BFGS formula*. In subsequent years, as the scale of tractable problems has increased exponentially, new ideas have also become important, including *automatic differentiation*, a technology that enables derivatives of computed functions to be determined automatically: the computer program itself is “differentiated,” so that as well as producing numerical outputs it also produces

their derivatives. The idea of automatic differentiation is an old one, but for various reasons, partly related to advances in sparse linear algebra and to the development of “reverse mode” formulations, it did not become fully practical until the work of Bischof, Carle, and Griewank in the 1990s.

Unconstrained optimization problems are relatively easy, but they are not typical; the true depth of this field is revealed by the methods that have been developed for dealing with constraints. Suppose a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is to be minimized subject to certain equality constraints $c_j(x) = 0$ and inequality constraints $d_j(x) \geq 0$, where $\{c_j\}$ and $\{d_j\}$ are also functions from \mathbb{R}^n to \mathbb{R} . Even the problem of stating local optimality conditions for solutions to such problems is nontrivial, a matter involving LAGRANGE MULTIPLIERS [III.66] and a distinction between active and inactive constraints. This problem was solved by what are now known as the *KKT conditions*, introduced by Kuhn and Tucker in 1951 and also twelve years earlier, it was subsequently realized, by Karush. Development of algorithms for constrained nonlinear optimization continues to be an active research topic today.

The problem of constraints brings us to the other strand of numerical optimization, linear programming. This subject was born in the 1930s and 1940s with Kantorovich in the Soviet Union and Dantzig in the United States. As an outgrowth of his work for the U.S. Air Force in the war, Dantzig invented in 1947 the famous SIMPLEX ALGORITHM [III.86] for solving linear programs. A linear program is nothing more than a problem of minimizing a linear function of n variables subject to m linear equality and/or inequality constraints. How can this be a challenge? One answer is that m and n may be large. Large-scale problems may arise through discretization of continuous problems and also in their own right. A famous early example was Leontiev's theory of input-output models in economics, which won him the Nobel Prize in 1973. Even in the 1970s the Soviet Union used an input-output computer model involving thousands of variables as a tool for planning the economy.

PUP: 'Even in' must stay according to Tim.

The simplex algorithm made medium- and large-scale linear programming problems tractable. Such a problem is defined by its *objective function*, the function $f(x)$ to be minimized, and its *feasible region*, the set of vectors $x \in \mathbb{R}^n$ that satisfy all the constraints. For a linear program the feasible region is a polyhedron, a closed domain bounded by hyperplanes, and the optimal value of f is guaranteed to be attained at one of

the vertex points. (A point is called a *vertex* if it is the unique solution of some subset of the equations that define the constraints.) The simplex algorithm proceeds by moving systematically downhill from one vertex to another until an optimal point is reached. All of the iterates lie on the boundary of the feasible region.

In 1984, an upheaval occurred in this field, triggered by Narendra Karmarkar at AT&T Bell Laboratories. Karmarkar showed that one could sometimes do much better than the simplex algorithm by working in the interior of the feasible region instead. Once a connection was shown between Karmarkar's method and the logarithmic barrier methods popularized by Fiacco and McCormick in the 1960s, new interior methods for linear programming were devised by applying techniques previously viewed as suitable only for nonlinear problems. The crucial idea of working in tandem with a pair of primal and dual problems led to today's powerful primal-dual methods, which can solve continuous optimization problems with millions of variables and constraints. Starting with Karmarkar's work, not only has the field of linear programming changed completely, but the linear and nonlinear sides of optimization are seen today as closely related rather than essentially different.

7 The Future

Numerical analysis sprang from mathematics; then it spawned the field of computer science. When universities began to found computer science departments in the 1960s, numerical analysts were often in the lead. Now, two generations later, most of them are to be found in mathematics departments. What happened? A part of the answer is that numerical analysts deal with continuous mathematical problems, whereas computer scientists prefer discrete ones, and it is remarkable how wide a gap this can be.

Nevertheless, the computer science side of numerical analysis is of crucial importance, and I would like to end with a prediction that emphasizes this aspect of the subject. Traditionally one might think of a numerical algorithm as a cut-and-dried procedure, a loop of some kind to be executed until a well-defined termination criterion is satisfied. For some computations this picture is accurate. On the other hand, beginning with the work of de Boor, Lyness, Rice and others in the 1960s, a less deterministic kind of numerical computing began to appear: *adaptive algorithms*. In an adaptive quadrature program of the simplest kind, two estimates of

the integral are calculated on each portion of a certain mesh and then compared to produce an estimate of the local error. Based on this estimate, the mesh may then be refined locally to improve the accuracy. This process is carried out iteratively until a final answer is obtained that aims to be accurate to a tolerance specified in advance by the user. Most such computations come with no guarantee of accuracy, but an exciting ongoing development is the advance of more sophisticated techniques of a posteriori error control that sometimes do provide guarantees. When these are combined with techniques of *interval arithmetic*, there is even the prospect of accuracy guaranteed with respect to rounding as well as discretization error.

First, computer programs for quadrature became adaptive; then programs for ODEs did as well. For PDEs, the move to adaptive programs is happening on a longer timescale. More recently there have been related developments in the computation of Fourier transforms, optimization, and large-scale numerical linear algebra, and some of the new algorithms adapt to the computer architecture as well as the mathematical problem. In a world where several algorithms are known for solving every problem, we increasingly find that the most robust computer program will be one that has diverse capabilities at its disposal and deploys them adaptively on the fly. In other words, numerical computation is increasingly embedded in intelligent control loops. I believe this process will continue, just as has happened in so many other areas of technology, removing scientists further from the details of their computations but offering steadily growing power in exchange. I expect that most of the numerical computer programs of 2050 will be 99% intelligent "wrapper" and just 1% actual "algorithm," if such a distinction makes sense. Hardly anyone will know how they work, but they will be extraordinarily powerful and reliable, and will often deliver results of guaranteed accuracy.

This story will have a mathematical corollary. One of the fundamental distinctions in mathematics is between linear problems, which can be solved in one step, and nonlinear ones, which usually require iteration. A related distinction is between forward problems (one step) and inverse problems (iteration). As numerical algorithms are increasingly embedded in intelligent control loops, almost every problem will be handled by iteration, regardless of its philosophical status. Problems of algebra will be solved by methods of analysis; and between linear and nonlinear, or forward and inverse, the distinctions will fade.

Table 1 Some algorithmic developments in the history of numerical analysis.

Year	Development	Key early figures
263	Gaussian elimination	Liu, Lagrange, Gauss, Jacobi
1671	Newton's method	Newton, Raphson, Simpson
1795	Least-squares fitting	Gauss, Legendre
1814	Gauss quadrature	Gauss, Jacobi, Christoffel, Stieltjes
1855	Adams ODE formulas	Euler, Adams, Bashforth
1895	Runge–Kutta ODE formulas	Runge, Heun, Kutta
1910	Finite differences for PDE	Richardson, Southwell, Courant, von Neumann, Lax
1936	Floating-point arithmetic	Torres y Quevedo, Zuse, Turing
1943	Finite elements for PDE	Courant, Feng, Argyris, Clough
1946	Splines	Schoenberg, de Casteljau, Bezier, de Boor
1947	Monte Carlo simulation	Ulam, von Neumann, Metropolis
1947	Simplex algorithm	Kantorovich, Dantzig
1952	Lanczos and conjugate gradient iterations	Lanczos, Hestenes, Stiefel
1952	Stiff ODE solvers	Curtiss, Hirschfelder, Dahlquist, Gear
1954	Fortran	Backus
1958	Orthogonal linear algebra	Aitken, Givens, Householder, Wilkinson, Golub
1959	Quasi-Newton iterations	Davidon, Fletcher, Powell, Broyden
1961	QR algorithm for eigenvalues	Rutishauser, Kublanovskaya, Francis, Wilkinson
1965	Fast Fourier transform	Gauss, Cooley, Tukey, Sande
1971	Spectral methods for PDE	Chebyshev, Lanczos, Clenshaw, Orszag, Gottlieb
1971	Radial basis functions	Hardy, Askey, Duchon, Micchelli
1973	Multigrid iterations	Fedorenko, Bakhvalov, Brandt, Hackbusch
1976	EISPACK, LINPACK, LAPACK	Moler, Stewart, Smith, Dongarra, Demmel, Bai
1976	Nonsymmetric Krylov iterations	Vinsome, Saad, van der Vorst, Sorensen
1977	Preconditioned matrix iterations	van der Vorst, Meijerink
1977	MATLAB	Moler
1977	IEEE arithmetic	Kahan
1982	Wavelets	Morlet, Grossmann, Meyer, Daubechies
1984	Interior methods in optimization	Fiacco, McCormick, Karmarkar, Megiddo
1987	Fast multipole method	Rokhlin, Greengard
1991	Automatic differentiation	Iri, Bischof, Carle, Griewank

8 Appendix: Some Major Numerical Algorithms

The list in [table 1](#) attempts to identify some of the most significant algorithmic (as opposed to theoretical) developments in the history of numerical analysis. In each case some of the key early figures are cited, more or less chronologically, and a key early date is given. Of course, any brief sketch of history like this must be an oversimplification. Distressing omissions of names occur throughout the list, including many early contributors in fields such as finite elements, preconditioning, and automatic differentiation, as well as more than half of the authors of the EISPACK, LINPACK, and LAPACK libraries. Even the dates can be questioned; the fast Fourier transform is listed as 1965, for example, since that is the year of the paper that brought it to

the world's attention, though Gauss made the same discovery 160 years earlier. Nor should one imagine that the years from 1991 to the present have been a blank! No doubt in the future we shall identify developments from this period that deserve a place in the table.

Further Reading

Ciarlet, P. G. 1978. *The Finite Element Method for Elliptic Problems*. Amsterdam: North-Holland.
Golub, G. H., and C. F. Van Loan. 1996. *Matrix Computations*, 3rd edn. Baltimore, MD: Johns Hopkins University Press.
Hairer, E., S. P. Nørsett (for volume I), and G. Wanner. 1993, 1996. *Solving Ordinary Differential Equations*, volumes I and II. New York: Springer.
Iserles, A., ed. 1992–. *Acta Numerica* (annual volumes). Cambridge: Cambridge University Press.
Nocedal, J., and S. J. Wright. 1999. *Numerical Optimization*. New York: Springer.

PUP: the fact that the date for 'Gaussian elimination' in the table predates Gauss is indeed OK. T&T note: check position of table before CRC. (Cannot appear earlier than page facing this reference.)

- Powell, M. J. D. 1981. *Approximation Theory and Methods*. Cambridge: Cambridge University Press.
- Richtmyer, R. D., and K. W. Morton. 1967. *Difference Methods for Initial-Value Problems*. New York: Wiley Interscience.

IV.22 Set Theory

Joan Bagaria

1 Introduction

Among all mathematical disciplines, set theory occupies a special place because it plays two very different roles at the same time: on the one hand, it is an area of mathematics devoted to the study of abstract sets and their properties; on the other, it provides mathematics with its foundation. This second aspect of set theory gives it philosophical as well as mathematical significance. We shall discuss both aspects of the subject in this article.

2 The Theory of Transfinite Numbers

Set theory began with the work of CANTOR [VI.54]. In 1874 he proved that there are more real numbers than there are algebraic ones, thus showing that infinite sets can be of different sizes. This also provided a new proof of the existence of TRANSCENDENTAL NUMBERS [III.43]. Recall that a real number is called *algebraic* if it is the solution of some polynomial equation

$$a_n X^n + a_{n-1} X^{n-1} + \cdots + a_1 X + a_0 = 0,$$

where the coefficients a_n are integers (and $a_n \neq 0$). Thus, numbers like $\sqrt{2}$, $\frac{3}{4}$, and the golden ratio, $\frac{1}{2}(1 + \sqrt{5})$, are algebraic. A transcendental number is one that is not algebraic.

What does it mean to say that there are “more” real numbers than algebraic ones, when there are infinitely many of both? Cantor defined two sets A and B to have the same size, or *cardinality*, if there is a bijection between them: that is, if there is a one-to-one correspondence between the elements of A and the elements of B . If there is no bijection between A and B , but there is a bijection between A and a *subset* of B , then A is of *smaller cardinality* than B . So what Cantor in fact showed was that the set of algebraic numbers had smaller cardinality than that of all real numbers.

In particular, Cantor distinguished between two different kinds of infinite set: COUNTABLE AND UNCOUNTABLE [III.11]. A countable set is one that can be put into one-to-one correspondence with the natural numbers.

In other words, it is a set that we can “enumerate,” assigning a different natural number to each of its elements. Let us see how this can be done for the algebraic numbers. Given a polynomial equation as above, let the number

$$|a_n| + |a_{n-1}| + \cdots + |a_0| + n$$

be called its *index*. It is easy to see that for every $k > 0$ there are only a finite number of equations of index k . For instance, there are only four equations of index 3 with strictly positive a_n , namely, $X^2 = 0$, $2X = 0$, $X + 1 = 0$, and $X - 1 = 0$, which have as solutions 0, -1 , and 1. Thus, we can enumerate the algebraic numbers by first enumerating all solutions of equations of index 1, then all solutions of equations of index 2 that we have not already enumerated, and so on. Therefore, the algebraic numbers are countable. Note that from this proof we also see that the sets \mathbb{Z} and \mathbb{Q} are countable.

Cantor discovered that, surprisingly, the set \mathbb{R} of real numbers is not countable. Here is Cantor's original proof. Suppose, aiming for a contradiction, that r_0, r_1, r_2, \dots is an enumeration of \mathbb{R} . Let $a_0 = r_0$. Choose the least k such that $a_0 < r_k$ and put $b_0 = r_k$. Given a_n and b_n , choose the least l such that $a_n < r_l < b_n$, and put $a_{n+1} = r_l$. And choose the least m such that $a_{n+1} < r_m < b_n$, and put $b_{n+1} = r_m$. Thus, we have $a_0 < a_1 < a_2 < \cdots < b_2 < b_1 < b_0$. Now let a be the limit of the a_n . Then a is a real number different from r_n , for all n , contradicting our assumption that the sequence r_0, r_1, r_2, \dots enumerates all real numbers.

Thus it was established for the first time that there are at least two genuinely different kinds of infinite sets. Cantor also showed that there are bijections between any two of the sets \mathbb{R}^n , $n \geq 1$, and even $\mathbb{R}^{\mathbb{N}}$, the set of all infinite sequences r_0, r_1, r_2, \dots of real numbers, so all these sets have the same (uncountable) cardinality.

From 1879 to 1884 Cantor published a series of works that constitute the origin of set theory. An important concept that he introduced was that of infinite, or “transfinite,” *ordinals*. When we use the natural numbers to count a collection of objects, we assign a number to each object, starting with 1, continuing with 2, 3, etc., and stopping when we have counted each object exactly once. When this process is over we have done two things. The more obvious one is that we have obtained a number n , the last number in the sequence, that tells us how many objects there are in the collection. But that is not all we have done: as we count we

also define an *ordering* on the objects that we were counting, namely the order in which we count them. This reflects two different ways in which we can think about the set $\{1, 2, \dots, n\}$. Sometimes all we care about is its size. Then, if we have a set X in one-to-one correspondence with $\{1, 2, \dots, n\}$, we conclude that X has cardinality n . But sometimes we also take note of the natural ordering on the set $\{1, 2, \dots, n\}$, in which case we observe that our one-to-one correspondence provides us with an ordering on X too. If we adopt the first point of view, then we are regarding n as a *cardinal*, and if we adopt the second, then we are regarding it as an ordinal.

If we have a countably infinite set, then we can think of that from the ordinal point of view too. For instance, if we define a one-to-one correspondence between \mathbb{N} and \mathbb{Z} by taking $0, 1, 2, 3, 4, 5, 6, 7, \dots$ to $0, 1, -1, 2, -2, 3, -3, \dots$, then we have not only shown that \mathbb{N} and \mathbb{Z} have the same cardinality, but also used the obvious ordering on \mathbb{N} to define an ordering on \mathbb{Z} .

Suppose now that we want to count the points in the unit interval $[0, 1]$. Cantor's argument given above shows that no matter how we assign numbers in this interval to the numbers $0, 1, 2, 3$, etc., we will run out of natural numbers before we have counted all points. However, when this happens, nothing prevents us from simply setting aside the numbers we have already counted and starting again. This is where transfinite ordinals come in: they are a continuation of the sequence $0, 1, 2, 3, \dots$ "beyond infinity," and they can be used to count bigger infinite sets.

To start with, we need an ordinal number that represents the first position in the sequence that comes straight after all the natural numbers. This is the first infinite ordinal number, which Cantor denoted by ω . In other words, after $0, 1, 2, 3, \dots$ comes ω . The ordinal ω has a different character from the previous ordinals, because although it has predecessors, it has no immediate predecessor (unlike 7 , say, which has immediate predecessor 6). We say that ω is a *limit ordinal*. But once we have ω , we can continue the ordinal sequence in a very simple way, just by adding 1 repeatedly. Thus, the sequence of ordinal numbers begins as follows:

$$0, 1, 2, 3, 4, 5, 6, 7, \dots, \omega, \omega + 1, \omega + 2, \omega + 3, \dots$$

After this comes the next limit ordinal, which it seems natural to call $\omega + \omega$, and which we can write as $\omega \cdot 2$. The sequence continues as

$$\omega \cdot 2, \omega \cdot 2 + 1, \omega \cdot 2 + 2, \dots, \omega \cdot n, \dots, \omega \cdot n + m, \dots$$

As this discussion indicates, there are two basic rules for generating new ordinals: adding 1 and passing to the limit. What we mean by "passing to the limit" is "assigning a new ordinal number to the position in the ordinal sequence that comes straight after all the ordinals obtained so far." For example, after all the ordinals $\omega \cdot n + m$ comes the next limit ordinal, which we write as $\omega \cdot \omega$, or ω^2 , and we obtain

$$\omega^2, \omega^2 + 1, \dots, \omega^2 + \omega, \dots, \omega^2 + \omega \cdot n, \dots, \omega^2 \cdot n, \dots$$

Eventually, we reach ω^3 and the sequence continues as

$$\omega^3, \omega^3 + 1, \dots, \omega^3 + \omega, \dots, \omega^3 + \omega^2, \dots, \omega^3 \cdot n, \dots$$

The next limit ordinal is ω^4 , and so on. The first limit ordinal after all the ω^n is ω^ω . And after $\omega^\omega, \omega^{\omega^\omega}, \omega^{\omega^{\omega^\omega}}, \dots$ comes the limit ordinal denoted by ε_0 . And on and on it goes.

In set theory, one likes to regard all mathematical objects as sets. For ordinals this can be done in a particularly simple way: we represent 0 by the empty set, and the ordinal number α is then identified with the set of all its predecessors. For instance, the natural number n is identified with the set $\{0, 1, \dots, n-1\}$ (which has cardinality n) and the ordinal $\omega + 3$ is identified with the set $\{0, 1, 2, 3, \dots, \omega, \omega + 1, \omega + 2\}$. If we think of ordinals in this way, then the ordering on the set of ordinals becomes set membership: if α comes before β in the ordinal sequence, then α is one of the predecessors of β and therefore an element of β . A critically important property of this ordering is that each ordinal is a *well-ordered set*, which means that every nonempty subset of it has a least element.

As we said earlier, cardinal numbers are used for measuring the sizes of sets, while ordinal numbers indicate the position in an ordered sequence. This distinction is much more apparent for infinite numbers than for finite ones, because then it is possible for two different ordinals to have the same size. For example, the ordinals ω and $\omega + 1$ are different but the corresponding sets $\{0, 1, 2, \dots\}$ and $\{0, 1, 2, \dots, \omega\}$ have the same cardinality, as figure 1 shows. In fact, all sets that can be counted using the infinite ordinals we have described so far are countable. So in what sense are different ordinals different? The point is that although two sets such as $\{0, 1, 2, \dots\}$ and $\{0, 1, 2, \dots, \omega\}$ have the same cardinality, they are not *order isomorphic*: that is, you cannot find a bijection ϕ from one set to the other such that $\phi(x) < \phi(y)$ whenever $x < y$. Thus, they are the same "as sets" but not "as ordered sets."

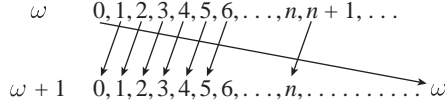


Figure 1 ω and $\omega + 1$ have the same cardinality.

Informally, the cardinal numbers are the possible sizes of sets. A convenient formal definition of a cardinal number is that it is an ordinal number that is bigger than all its predecessors. An important example of such an ordinal is the set of all countable ordinals, which Cantor denoted by ω_1 . This is the first uncountable ordinal: uncountable since it cannot include itself as an element, and the first one because all its elements are countable. (If this seems paradoxical, consider the ordinal ω : it is infinite, but all its elements are finite.) Therefore, it is also a cardinal number, and when we consider this aspect of it rather than its order structure we call it \aleph_1 , again following Cantor.

This process can be repeated. The set of all ordinals of cardinality \aleph_1 (or equivalently the set of all ordinals that can be put in one-to-one correspondence with the first uncountable ordinal ω_1) is the smallest ordinal that has cardinality greater than \aleph_1 . As an ordinal it is called ω_2 and as a cardinal it is called \aleph_2 . We can continue, generating a whole sequence of ordinals $\omega_1, \omega_2, \omega_3, \dots$ of larger and larger cardinality. Moreover, using limits as well, we can continue this sequence transfinitely: for example, the ordinal ω_ω is the limit of all the ordinals ω_n . As we do this, we also produce the sequence of infinite, or transfinite, cardinals:

$$\aleph_0, \aleph_1, \dots, \aleph_\omega, \aleph_{\omega+1}, \dots, \aleph_{\omega^\omega}, \dots, \\ \aleph_{\omega_1}, \dots, \aleph_{\omega_2}, \dots, \aleph_{\omega_\omega}, \dots$$

Given two natural numbers, we can calculate their sum and product. A convenient set-theoretic way to define these binary operations is as follows. Given two natural numbers m and n , take any two disjoint sets A and B of size m and n , respectively; $m + n$ is then the size of the union $A \cup B$. As for the product, it is the size of the set $A \times B$, the set of all ordered pairs (a, b) with $a \in A$ and $b \in B$. (For this set, which is called the *Cartesian product*, we do not need A and B to be disjoint.)

The point of these definitions is that they apply just as well to infinite cardinal numbers: just replace m and n in the above definitions by two infinite cardinals κ and λ . The resulting arithmetic of transfinite cardinals is very simple, however. It turns out that for all

transfinite cardinals \aleph_α and \aleph_β ,

$$\aleph_\alpha + \aleph_\beta = \aleph_\alpha \aleph_\beta = \max(\aleph_\alpha, \aleph_\beta) = \aleph_{\max(\alpha, \beta)}.$$

However, it is also possible to define cardinal exponentiation, and for this the picture changes completely. If κ and λ are two cardinals, then κ^λ is defined as the cardinality of the Cartesian product of λ copies of any set of cardinality κ . Equivalently, it is the cardinality of the set of all functions from a set of cardinality λ into a set of cardinality κ . Again, if κ and λ are *finite* numbers, this gives us the usual definition: for instance, the number of functions from a set of size 3 to a set of size 4 is 4^3 . What happens if we take the simplest nontrivial transfinite example, 2^{\aleph_0} ? Not only is this question extremely hard, there is a sense in which it cannot be resolved, as we shall see later.

The most obvious set of cardinality 2^{\aleph_0} is the set of functions from \mathbb{N} to the set $\{0, 1\}$. If f is such a function, then we can regard it as giving the binary expansion of the number

$$x = \sum_{n \in \mathbb{N}} f(n) 2^{-(n+1)},$$

which belongs to the closed interval $[0, 1]$. (The power is $2^{-(n+1)}$ rather than 2^{-n} because we are using the convention, standard in set theory, that 0 is the first natural number rather than 1.) Since every point in $[0, 1]$ has at most two different binary representations, it follows easily that 2^{\aleph_0} is also the cardinality of $[0, 1]$, and therefore also the cardinality of \mathbb{R} . Thus, 2^{\aleph_0} is uncountable, which means that it is greater than or equal to \aleph_1 . Cantor conjectured that it is exactly \aleph_1 . This is the famous *continuum hypothesis*, which will be discussed at length in section 5 below.

It is not immediately obvious, but there are many mathematical contexts in which transfinite ordinals occur naturally. Cantor himself devised his theory of transfinite ordinals and cardinals as a result of his attempts, which were eventually successful, to prove the continuum hypothesis for closed sets. He first defined the *derivative* of a set X of real numbers to be the set you obtain when you throw out all the “isolated” points of X . These are points x for which you can find a small neighborhood around x that contains no other points in X . For example, if X is the set $\{0\} \cup \{1, \frac{1}{2}, \frac{1}{3}, \dots\}$, then all points in X are isolated except for 0, so the derivative of X is the set $\{0\}$.

In general, given a set X , we can take its derivative repeatedly. If we set $X^0 = X$, then we obtain a sequence $X^0 \supseteq X^1 \supseteq X^2 \supseteq \dots$, where X^{n+1} is the derivative of X^n . But the sequence does not stop here: we can

take the intersection of all the X^n and call it X^ω , and if we do that, then we can define $X^{\omega+1}$ to be the derivative of X^ω , and so on. Thus, the reason that ordinals appear naturally is that we have *two* operations, taking the derivative and taking the intersection of everything so far, which correspond to successors and limits in the ordinal sequence. Cantor initially regarded superscripts such as $\omega + 1$ as “tags” that marked the transfinite stages of the derivation. These tags later became the countable ordinal numbers.

Cantor proved that for every closed set X there must be a countable ordinal α (which could be finite) such that $X^\alpha = X^{\alpha+1}$. It is easy to show that each X^β in the sequence of derivatives is closed, and that it contains all but countably many points of the original set X . Therefore, X^α is a closed set that contains no isolated points. Such sets are called *perfect sets* and it is not too hard to show that they are either empty or have cardinality 2^{\aleph_0} . From this it follows that X is either countable or of cardinality 2^{\aleph_0} .

The intimate connection, discovered by Cantor, between transfinite ordinals and cardinals and the structure of the continuum was destined to leave its mark on the entire subsequent development of set theory.

3 The Universe of All Sets

In the discussion so far we have taken for granted that every set has a cardinality, or in other words that for every set X there is a unique cardinal number that can be put into one-to-one correspondence with X . If κ is such a cardinal and $f : X \rightarrow \kappa$ is a bijection (recall that we identify κ with the set of all its predecessors), then we can define an ordering on X by taking $x < y$ if and only if $f(x) < f(y)$. Since κ is a well-ordered set, this makes X into a well-ordered set. But it is far from obvious that every set can be given a well-ordering: indeed, it is not obvious even for the set \mathbb{R} . (If you need convincing of this, then try to find one.)

Thus, to make full use of the theory of transfinite ordinals and cardinals and to solve some of the fundamental problems—such as computing where in the aleph hierarchy of infinite cardinals the cardinal of \mathbb{R} is—one must appeal to the *well-ordering principle*: the assertion that every set can be well-ordered. Without this assertion, one cannot even make sense of the questions. The well-ordering principle was introduced by Cantor, but he was unable to prove it. HILBERT [VI.63] listed proving that \mathbb{R} could be well-ordered as part of the first problem in his celebrated list of twenty-three

unsolved mathematical problems presented in 1900 at the Second International Congress of Mathematicians in Paris. Four years later, Ernst Zermelo gave a proof of the well-ordering principle that drew a lot of criticism for its use of THE AXIOM OF CHOICE [III.1] (AC), a principle that had been tacitly used for many years but which was now brought into focus by Zermelo's result. AC states that *for every set X of pairwise-disjoint nonempty sets there is a set that contains exactly one element from each set in X* . In a second, much more detailed, proof published in 1908, Zermelo spells out some of the principles or axioms involved in his proof of the well-ordering principle, including AC.

In that same year, Zermelo published the first axiomatization of set theory, the main motivation being the need to continue with the development of set theory while avoiding the logical traps, or paradoxes, that originated in the careless use of the intuitive notion of a set (see THE CRISIS IN THE FOUNDATIONS OF MATHEMATICS [II.7]). For instance, it seems intuitively clear that every property determines a set, namely, the set of those objects that have that property. But then consider the property of *being an ordinal number*. If this property determined a set, this would be the set of all ordinal numbers. But a moment of reflection shows that there cannot be such a set, since it would be well-ordered and would therefore correspond to an ordinal greater than all ordinals, which is absurd. Similarly, the property of *being a set that is not an element of itself* cannot determine a set, for otherwise we fall into Russell's paradox, that if A is such a set, then A is an element of A if and only if A is not an element of A , which is absurd. Thus, not every collection of objects, not even those that are defined by some property, can be taken to be a set. So what is a set? Zermelo's 1908 axiomatization provides the first attempt to capture our intuitive notion of set in a short list of basic principles. It was later improved through contributions from SKOLEM [VI.81], Abraham Fraenkel, and VON NEUMANN [VI.91], becoming what is now known as *Zermelo-Fraenkel set theory with the axiom of choice*, or ZFC.

The basic idea behind the axioms of ZFC is that there is a “universe of all sets” that we would like to understand, and the axioms give us the tools we need to build sets out of other sets. In usual mathematical practice we take sets of integers, sets of real numbers, sets of functions, etc., but also sets of sets (such as sets of open sets in a TOPOLOGICAL SPACE [III.92]), sets of sets of sets (such as sets of open covers), and so on. Thus, the universe of all sets should consist not only of sets

PUP: Tim prefers
'but' to 'and' here.

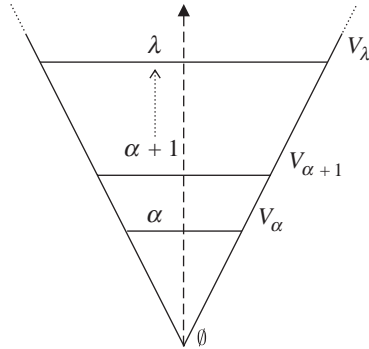


Figure 2 The universe V of all pure sets.

of objects, but also of sets of sets of objects, etc. Now it turns out that it is much more convenient to dispense with “objects” altogether and consider only sets whose elements are sets, whose elements are also sets, etc. Let us call those sets “pure sets.” The restriction to pure sets is technically advantageous and yields a more elegant theory. Moreover, it is possible to model traditional mathematical concepts such as real numbers using pure sets, so one does not lose any mathematical power. Pure sets are built from nothing, i.e., the empty set, by successively applying the “set of” operation. A simple example is $\{\emptyset, \{\emptyset, \{\emptyset\}\}\}$: to build this we start by forming $\{\emptyset\}$, then $\{\emptyset, \{\emptyset\}\}$, and putting these two sets together gives us $\{\emptyset, \{\emptyset, \{\emptyset\}\}\}$. Thus, at every stage we form all the sets whose elements are sets already obtained in the previous stages. Once again, this can be continued transfinitely: at limit stages we collect into a set all the sets obtained so far, and keep going. The universe of all (pure) sets, represented by the letter V and usually drawn as a V-shape with a vertical axis representing the ordinals (see figure 2), therefore forms a cumulative well-ordered hierarchy, indexed by the ordinal numbers, beginning with the empty set \emptyset . That is, we let

$$\begin{aligned} V_0 &= \emptyset, \\ V_{\alpha+1} &= \mathcal{P}(V_\alpha), \quad \text{the set of all subsets of } V_\alpha, \\ V_\lambda &= \bigcup_{\beta < \lambda} V_\beta, \quad \text{the union of all the } V_\beta, \beta < \lambda, \end{aligned}$$

if λ is a limit ordinal.

The universe of all sets is then the union of all the sets V_α such that α is an ordinal. More concisely,

$$V = \bigcup_{\alpha} V_\alpha.$$

3.1 The Axioms of ZFC

The ZFC axioms, stated informally, are the following.

- (i) **Extensionality.** If two sets have the same elements, they are equal.
- (ii) **Power set.** For every set x there is a set $\mathcal{P}(x)$ whose elements are all the subsets of x .
- (iii) **Infinity.** There is an infinite set.
- (iv) **Replacement.** If x is a set and ϕ is a *function-class*¹ restricted to x , then there is a set $y = \{\phi(u) : u \in x\}$.
- (v) **Union.** For every set x , there is a set $\bigcup x$ whose elements are all the elements of the elements of x .
- (vi) **Regularity.** Every set x belongs to V_α , for some ordinal α .
- (vii) **Axiom of choice (AC).** For every set X of pairwise-disjoint nonempty sets there is a set that contains exactly one element from each set in X .

Usually a further axiom appears on this list, called the *pairing axiom*. It asserts that for any two sets A and B the set $\{A, B\}$ exists. In particular, $\{A\}$ exists. Applying the union axiom to the set $\{A, B\}$ one then gets the union $A \cup B$ of A and B . But pairing can be derived from the other axioms. Another important axiom that appeared in Zermelo’s original list, one that is both natural and very useful, is the *axiom of separation*. It states that for every set A and every *definable property* P , the set of elements of A that have the property P is also a set. But this axiom is a consequence of the axiom of replacement, so there is no need to include it in the list. Using the axiom of separation one can easily prove the existence of the empty set \emptyset , as well as the intersection $A \cap B$ and difference $A - B$ of any two sets A and B . The axiom of regularity is also known as the *axiom of foundation* and it is usually stated as follows: every nonempty set X has an \in -minimal element, i.e., an element that no element of X belongs to. In the presence of the other axioms the two formulations are equivalent. We chose the formulation in terms of the V_α s to stress the fact that this is a natural axiom based on the construction of the universe of all sets. But it is important to notice that the notions of “ordinal” and the “cumulative hierarchy of V_α s” need not appear in the formulation of the axioms of ZFC.

The axioms of ZFC lead a kind of double life. On the one hand, they tell us the things we can do with

1. A function-class can be thought of as a function that is given as a definition rather than an object that has to exist as a set. The concept will be made precise in section 3.2.

sets. In this sense, ZFC is just like any other collection of axioms for algebraic structures, e.g., the axioms for GROUPS [I.3 §2.1], or FIELDS [I.3 §2.2]: in both cases they give rules for creating new objects from old ones, though there are more rules for sets than there are for group or field elements and they are more complicated. Thus, just as one studies abstract groups, i.e., algebraic structures that satisfy the axioms for groups, so one can study the mathematical structures that satisfy the axioms of ZFC. These are called *models of ZFC*. Since, for reasons to be explained below, models of ZFC are not easy to come by, one is also interested in models of fragments of ZFC: that is, of axiom systems A that consist of just some of the axioms of ZFC. A model of a fragment A of ZFC is defined to be a pair $\langle M, E \rangle$, where M is a nonempty set and E is a binary relation on M , such that all axioms of A are true when the elements of M are interpreted as the sets and E is interpreted as the membership relation. For example, if A includes the union axiom, then for every element x of M there must be an element y of M such that xEy if and only if there exists w such that zEw and wEx . (If we replaced E by \in and “element of M ” by “set” in the last sentence, then we would recover the usual union axiom.)

The set $\langle V_\omega, \in \rangle$ is a model of all the axioms of ZFC except infinity, and $\langle V_{\omega+\omega}, \in \rangle$ is a model of ZFC except replacement. (To see why replacement fails, let x be the set ω and define a function ϕ on x by letting $\phi(n)$ equal $\omega + n$. The range of ϕ belongs to $V_{\omega+\omega+1}$ but not to $V_{\omega+\omega}$ because the ordinal $\omega + \omega$ does not belong to any set $V_{\omega+n}$ and $V_{\omega+\omega}$ is the union of the sets $V_{\omega+n}$.) For both these models, we took E to be \in , but one can also look at a completely different relation E on a set M , and see whether it happens to satisfy some of the axioms of ZFC. For example, take the pair $\langle \mathbb{N}, E \rangle$, where mEn if and only if the m th digit (counting from right to left) in the binary expansion of n is 1. This is a model of ZFC without the axiom of infinity, as the reader may care to check.

The other way of thinking of the ZFC axioms is that they tell us how to build up the hierarchy of the V_α s. Axiom (i), the axiom of extensionality, states that a set is something entirely determined by its elements. Axioms (ii)–(v) are tailored to construct V . The power-set axiom is what we use to get from V_α to $V_{\alpha+1}$. The axiom of infinity allows the construction to go into the transfinite. Indeed, in the context of the other ZFC axioms, this axiom is equivalent to the assertion that ω exists. The axiom of replacement is used to continue the construction of V at limit stages λ . To see this, con-

sider the function defined by $F(x) = y$ if and only if x is an ordinal and $y = V_x$. The range of F restricted to λ then consists of all V_β with $\beta < \lambda$. By the axiom of replacement these sets form a set. Now, by an application of the union axiom to this set one obtains V_λ . Finally, the axiom of regularity states that all sets are obtained in this way: that is, the universe of all sets is precisely V . This rules out pathologies, such as sets that belong to themselves. The point is that for every set X there is a first α such that $X \in V_{\alpha+1}$. This α is called the *rank* of X and it marks the stage of the cumulative hierarchy where X was formed. So X could not possibly be an element of itself, since all elements of X must have a rank strictly smaller than the rank of X . The axiom of choice is equivalent, in the context of the other ZFC axioms, to the well-ordering principle.

3.2 Formulas and Models

The ZFC axioms can be formalized using the language of *first-order logic for sets*. The symbols of first-order logic are *variables* such as x, y, z, \dots ; the *quantifiers* “ \forall ” (for all) and “ \exists ” (there exists); the *logical connectives* “ \neg ” (not), “ \wedge ” (and), “ \vee ” (or), “ \rightarrow ” (if ..., then ...), and “ \leftrightarrow ” (if and only if); the equality symbol “ $=$ ”; and parentheses. To make this first-order logic *for sets* we add one other symbol, “ \in ,” standing for “is an element of,” and the quantifiers are understood to range over sets. Here is how the axiom of extensionality is expressed in this language:

$$\forall x \forall y (\forall z (z \in x \leftrightarrow z \in y) \rightarrow x = y).$$

This reads as: for every set x and every set y , if every set z belongs to x if and only if it belongs to y (i.e., if x and y have the same elements), then x and y are equal. It is an example of a *formula* in our language. Formulas can be defined inductively as follows. The *atomic formulas* are $x = y$ and $x \in y$. Using quantifiers and logical connectives one can build up more complicated formulas using the following rules: if φ and ψ are formulas, then so are $\neg\varphi$, $(\varphi \wedge \psi)$, $(\varphi \vee \psi)$, $(\varphi \rightarrow \psi)$, $(\varphi \leftrightarrow \psi)$, $\forall x\varphi$, and $\exists x\varphi$. Thus, formulas are the formal counterpart of sentences in English (or in any other natural language) that talk only about sets and the membership relation. (For another discussion of formal languages, see LOGIC AND MODEL THEORY [IV.23 §1].)

Conversely, any formula of the formal language can be interpreted as a sentence (in English) about sets, and it makes sense to ask whether the interpreted sentence is true or not. Usually, by “true” we mean “true in the

universe V of all sets,” but it also makes sense to ask about the truth or falsity of a formula in any structure of the form $\langle M, E \rangle$, where E is a binary relation on M . For example, the formula $\forall x \exists y \ x \in y$ is true in all models $\langle M, E \rangle$ of ZFC, while the formula $\exists x \forall y \ y \in x$ is false (because of the axiom of regularity). Any formula that can be deduced from the axioms of ZFC is true in all models of ZFC.

Once we have defined what a formula is, we are in a position to make many statements precise that would otherwise not be. For example, the axiom of replacement involves the notion of a “function-class.” To make proper sense of it one formulates it in terms of first-order formulas. For example, the operation that takes each set a to the singleton $\{a\}$ is definable, and this depends on the fact that the statement $y = \{x\}$ can be expressed by the formula $\forall z (z \in y \leftrightarrow z = x)$. It is not a function, since it is defined on all sets, and the universe of all sets is not a set. This is why we use the different phrase “function-class.” In addition, we sometimes allow *parameters* in our definitions of function-classes. For example, the function-class that, for a fixed set b , takes each set a to the set $a \cap b$ is defined by the formula $\forall z (z \in y \leftrightarrow z \in x \wedge z \in b)$, which depends on the set b : we call b a parameter and we say that the function-class is *definable with parameters*. More generally, a function-class is a function on sets given by a formula. But the function itself may not exist as a set, since its domain may contain all sets, or all ordinals, etc. Since the axiom of replacement is a statement about all function-classes, it is not in fact a single axiom but rather an “axiom-scheme,” consisting of one axiom for each function-class.

An important consequence of the fact that ZFC can be formalized in first-order logic is that it is subject to a remarkable theorem of Löwenheim and Skolem. The Löwenheim-Skolem theorem is a general result about first-order formal languages; in the particular case of ZFC, it says that if ZFC has a model, then it has a countable model. More precisely, given any model $M = \langle M, E \rangle$ of ZFC, there is a model N of ZFC contained in M that is countable and that satisfies exactly the same sentences as M . At first, this may seem paradoxical, for how can ZFC have a countable model if one can prove in ZFC that there are uncountable sets? Does the theorem not lead to a contradiction and therefore imply that there are no models of ZFC? Not quite. Suppose that we have a countable model N of ZFC and a set a in N . If we want to show that the statement “ a is countable” is true in N , then we must show that *in N* there is a surjective map

from ω to a . But it is possible for such a map to exist in V , or in some model M that is larger than N , without existing in N , because V and M contain more sets, and therefore more functions, than N does. In such a case, a is uncountable from the point of view of N but countable from the point of view of M or V .

Far from presenting a problem, the relativity of certain set-theoretic notions, like being countable or having a certain cardinality, with respect to different models of ZFC is an important phenomenon which, even if a bit disconcerting at first, may be used to great advantage in consistency proofs (see section 5 below).

It is not difficult to see that all the axioms of ZFC are true in V , which is hardly surprising since they were designed for that to happen. But the ZFC axioms may conceivably hold in some smaller universes. That is, there may be some class M properly contained in V , or even some set M , and therefore by the Löwenheim-Skolem theorem also some countable set M , which is a model of ZFC. As we shall see, while the existence of models of ZFC cannot be proved in ZFC, the fact that one can consistently assume that they exist—provided ZFC is consistent, of course—is of the greatest importance for set theory.

4 Set Theory and the Foundation of Mathematics

As we have seen, we can use ZFC to develop the theory of transfinite numbers. But it turns out that all standard mathematical objects may be viewed as sets, and all classical mathematical theorems can be proved from ZFC using the usual logical rules of proof. For example, real numbers can be defined as certain sets of rational numbers, which can be defined as EQUIVALENCE CLASSES [I.2 §2.3] of ordered pairs of integers. The ordered pair (m, n) can be defined as the set $\{m, \{m, n\}\}$, integers can be defined as equivalence classes of ordered pairs of positive integers, and positive integers can be thought of as finite ordinals, which as we have seen can be defined as sets. Tracing back, one finds that a real number can be regarded as a set of sets of sets of sets of sets of finite ordinals. Similarly, all the usual mathematical objects—such as algebraic structures, vector spaces, topological spaces, smooth manifolds, dynamical systems, and so on—can be shown to exist in ZFC. Theorems concerning these objects can be expressed in the formal language of ZFC, as can their proofs. Of course, writing out a complete proof using the formal language

would be extremely laborious, and the result would not only be very long but also virtually impossible to understand. It is important, however, to convince oneself that in principle it can be done. It is the fact that all standard mathematics can be formulated and developed within the axiomatic system of ZFC that makes *metamathematics* possible, that is, the rigorous mathematical study of mathematics itself. For example, it allows us to think about whether a mathematical statement has a proof: once we have rigorous definitions of “mathematical statement” and “proof,” the question of whether a proof exists becomes a mathematical one with a determinate answer.

4.1 Undecidable Statements

In mathematics the truth of a mathematical statement φ is established by means of a proof from basic principles or axioms. Similarly, the falsity of φ is established by a proof of $\neg\varphi$. It is tempting to believe that there must always be a proof of either φ or $\neg\varphi$, but in 1931 GÖDEL [VI.92] proved in his famous INCOMPLETENESS THEOREMS [V.18] that this is not the case. The first incompleteness theorem says that in every axiomatic formal system that is consistent and rich enough to develop basic arithmetic there are *undecidable* statements: that is, statements such that neither they nor their negations are provable in the system. In particular, there are statements of the formal language of set theory that are neither provable nor disprovable from the ZFC axioms, supposing, that is, that ZFC is consistent.

But is ZFC consistent? The statement that asserts the consistency of ZFC, usually written as $\text{CON}(\text{ZFC})$, is the translation into the language of set theory of:

$0 = 1$ is not provable in ZFC.

This statement asserts that the sequence of symbols $0 = 1$ is not the last step of any formal proof from ZFC. One can encode a formal proof as a finite sequence of natural numbers that satisfies certain arithmetical properties, and thereby regard the above statement as an arithmetical one. Gödel’s second incompleteness theorem says that in any consistent axiomatic formal system that is rich enough to develop basic arithmetic, the arithmetical statement that asserts the consistency of the system cannot be proved. Thus, if ZFC is consistent, then its consistency can neither be proved nor disproved in ZFC.

ZFC is currently accepted as the standard formal system in which to develop mathematics. Thus, the truth

of a mathematical statement is firmly established if its translation into the language of set theory is provable in ZFC. But what about undecidable statements? Since ZFC embodies all standard mathematical methods, the fact that a given mathematical statement φ is undecidable in ZFC means that the truth or falsity of φ cannot be established by means of usual mathematical practice. If all undecidable statements were like $\text{CON}(\text{ZFC})$, this would probably not be a cause of worry, since they seem not to directly affect the kind of mathematical problems that people are usually interested in. But for better or worse this is not so. As we will see, there are many statements of mathematical interest that are undecidable in ZFC.

There is an obvious way of showing that a mathematical statement has a proof: you just find one. But how can it be possible to prove, mathematically, that a given mathematical statement φ is undecidable in ZFC? This question has a short but far-reaching answer. If we can find a model M of ZFC in which φ is false, then there cannot be a proof of φ (because that proof would show that φ was true in M). Therefore, if we can find models M and N of ZFC with φ true in M and false in N , we can conclude that φ is undecidable.

Unfortunately, a consequence of Gödel’s second incompleteness theorem is that it is not possible to prove in ZFC the existence of a model of ZFC. This is because another theorem of Gödel, called the *completeness theorem* for first-order logic, asserts that ZFC is consistent if and only if it has a model. However, we can get around this difficulty by splitting the proof of the undecidability of φ into two *relative consistency* proofs: the first is a proof that if ZFC is consistent, then so is ZFC plus φ ; and the second is a proof that if ZFC is consistent, then so is ZFC plus the negation of φ . That is, one assumes that there is a model M of ZFC and proves the existence of two models of ZFC: one where φ holds, and one where it fails. One can then conclude that either φ and its negation are both unprovable in ZFC, or ZFC is inconsistent, in which case everything is provable.

One of the most surprising results of twentieth-century mathematics is that the continuum hypothesis is undecidable in ZFC.

5 The Continuum Hypothesis

Cantor’s continuum hypothesis (CH), first formulated in 1878, states that every infinite set of real numbers is either countable or has the same cardinality as \mathbb{R} . In ZFC, since AC implies that every set, and in particular every infinite set of real numbers, can be put

into one-to-one and onto correspondence with a cardinal number, one can easily see that CH is equivalent to the assertion that the cardinality of \mathbb{R} is \aleph_1 , or equivalently, that $2^{\aleph_0} = \aleph_1$, the version of the statement that we mentioned earlier.

Solving CH was the first problem in Hilbert's famous list of twenty-three unsolved problems, and has been one of the main driving forces for the development of set theory. In spite of many attempts at proving CH by Cantor himself and by many leading mathematicians of the first third of the twentieth century, no major progress was made until, sixty years after its formulation, Gödel was able to prove its consistency with ZFC.

5.1 The Constructible Universe

In 1938, Gödel found a way to construct, starting with a model M of ZFC, another model of ZFC, contained in M , where CH holds. He thereby proved the relative consistency of CH with ZFC. Gödel's model is known as the *constructible universe* and is represented by the letter L . Since M is a model of ZFC, we may view M as the universe V of all sets. Then L is built inside M in a way that is similar to how we built V , but with the following important difference. When we passed from V_α to $V_{\alpha+1}$ we took *all* subsets of V_α , but to go from L_α to $L_{\alpha+1}$ one takes only those subsets of L_α that are *definable* in L_α . That is, $L_{\alpha+1}$ consists of all sets of the form $\{a : a \in L_\alpha \text{ and } \varphi(a) \text{ holds in } L_\alpha\}$, where $\varphi(x)$ is a formula of the language of set theory that may mention elements of L_α . If λ is a limit ordinal, then L_λ is just the union of all the L_α , $\alpha < \lambda$, and L is the union of all the L_α , α an ordinal. Of course, we can also build L inside V . This is the *real* L , the universe of all constructible sets.

One important observation is that to build L it is not necessary to use AC, and so we do not require AC to hold in M . But once L is constructed it can be verified that AC holds in L , as do the other axioms of ZFC. The verification of AC is based on the fact that every element of L is defined at some stage α , and so it is uniquely determined by a formula and some ordinals. Therefore, any sensible well-ordering of all the formulas will naturally yield a well-ordering of L , and thus of every set in L . This shows that if ZF (i.e., ZFC minus AC) is consistent, then so is ZFC. In other words, if we add AC to the ZF axioms, then no contradiction is introduced into the system. This is very reassuring, for although AC has many desirable consequences it also has some that at first sight can appear counterintuitive, such as THE BANACH-TARSKI PARADOX [V.3].

That CH holds in L is due to the fact that in L every real number appears at some countable stage of the construction, i.e., in some L_α , where α is countable in L . To prove this, one shows first that every real r belongs to some L_β that satisfies a finite number of axioms of ZFC that are sufficient to build L , where β is an ordinal that is not necessarily countable. Then, with the help of the Löwenheim-Skolem theorem, one can show that there is a countable subset X of L_β that contains r and satisfies the same axioms as L_β . And then one shows that X must be isomorphic to L_α for some countable ordinal α , via an isomorphism that is the identity on r ; this finishes the proof that r appears at a countable stage. But since there are only \aleph_1 countable ordinals, and L_α is countable for each countable ordinal α , there can be only \aleph_1 real numbers.

Since, for each ordinal α , L_α contains only the sets that are strictly necessary, namely those that were explicitly definable in one of the previous stages, L is the smallest possible model of ZFC containing all the ordinals, and in it the cardinality of \mathbb{R} is also the smallest possible, namely \aleph_1 . In fact, in L the *generalized continuum hypothesis* (GCH) holds: that is, for every ordinal α , 2^{\aleph_α} has the smallest possible value, namely, $\aleph_{\alpha+1}$.

The theory of constructible sets went through an extraordinary development in the hands of Ronald Jensen. He showed that in L a well-known conjecture called Suslin's hypothesis was false (see section 10 below) and isolated two important combinatorial principles, known as \diamond (diamond) and \square (square), that hold in L . These two principles, which will not be defined here, enable us to carry out constructions of uncountable mathematical structures by induction on the ordinals in such a way that the construction does not break down at limit stages. This is extremely useful, because it allows one to prove consistency results without going to the trouble of analyzing constructible sets: if you can deduce a statement φ from \diamond or \square , then it holds in L , since, by Jensen's results, \diamond and \square hold in L ; it follows that φ is consistent with ZFC.

There is also an important generalization of the notion of constructibility, called *inner model theory*. Given any set A it is possible to build the *constructible closure* of A , which is the smallest model of ZF that contains all ordinals and A . This model, called $L(A)$, is built in the same way as L , but instead of beginning with the empty set one begins with the *transitive closure* of A , which consists of A , the elements of A , the elements of the elements of A , and so on. Models of

this sort are examples of *inner models*: that is, models of ZF that contain all the ordinals and all the elements of their elements. Especially prominent are the inner models $L(r)$, where r is a real number, and $L(\mathbb{R})$, the constructible closure of the set of real numbers. Also very important are the inner models of large-cardinal axioms, which will be discussed in section 6 below.

After the result of Gödel, and given the repeated failed attempts to prove CH in ZFC, the idea started to take shape that maybe it was undecidable. To prove this, it was necessary to find a way to build a model of ZFC in which CH is false. This was finally accomplished twenty-five years later, in 1963, by Paul Cohen, using a revolutionary new technique called *forcing*.

5.2 Forcing

The forcing technique is an extremely flexible and powerful tool for building models of ZFC. It allows one to construct models with the most diverse properties and with great control over the statements that will hold in the model being constructed. It has made it possible to prove the consistency of many statements with ZFC that were not previously known to be consistent, and this has led to many undecidability results.

In a manner reminiscent of the way one passes from a field K to an algebraic extension $K[a]$, one goes from a model M of ZFC to a *forcing extension* $M[G]$ that is also a model of ZFC. However, the forcing method is far more complex, both conceptually and technically, involving set-theoretic, combinatorial, topological, logical, and metamathematical aspects.

To give an idea of how it works, let us consider Cohen's original problem of starting from a model M of ZFC and obtaining from it a model where CH fails. The only thing we know about M is that it is a model of ZFC, and as far as we know CH may hold in it. In fact, for all we know, M might be the constructible universe L : perhaps when we build L inside M we obtain the whole of M . Therefore, when we extend M we shall have to add to it some new real numbers to ensure that in the extension $M[G]$ there will be at least \aleph_2 of them. More precisely, we need the model $M[G]$ to satisfy the sentence that says that there are at least \aleph_2 -many real numbers. However, the "real numbers" in $M[G]$ may not be real numbers in the actual universe V : all that matters is that in $M[G]$ they satisfy sentences that say "I am a real number." Similarly, the element of $M[G]$ that plays the role of the cardinal \aleph_2 need not be the actual cardinal \aleph_2 in V .

In order to explain the method, let us consider the simpler problem of adding to M just a single new real number r . To make things even simpler, let us think of r as just the binary representation of a real in $[0, 1]$. In other words, r is an infinite binary sequence in the real world V .

A first difficulty is that M may already contain all infinite binary sequences, in which case we will not be able to find one to add. However, by the Löwenheim-Skolem theorem, every model M of ZFC has a countable submodel N that satisfies exactly the same sentences of the language of set theory as M . Let us emphasize that N is countable in the real world, that is, in V ; so there is, outside N , a function that enumerates all its elements. Nevertheless, N will contain sets x for which the sentence that says " x is uncountable" is true in N . Since M was a model of ZFC, so is N . So, since we really do not care about the size of M , but only that it is a model of ZFC, we may as well assume that $M = N$, so that M itself is countable. And now, since there are uncountably many infinite binary sequences, there are plenty of them that do not belong to M .

So, can we just pick any one of them and add it to M ? Well, no. The problem is that there are some binary sequences that have a great influence on any model that contains them. For example, we can encode any countable ordinal α as a real number as follows. First let f be a bijection from \mathbb{N} to α and define a subset $A \subset \mathbb{N}^2$ to be $\{(m, n) \in \mathbb{N}^2 : f(m) < f(n)\}$. Now choose a bijection g from \mathbb{N} to \mathbb{N}^2 and let $c(n) = 1$ if and only if $g(n) \in A$. If g is sufficiently explicit (as it can easily be chosen to be), then any model M that contains the infinite binary sequence c must contain the ordinal α , since α can be built out of c using the axioms of ZFC.

To see why this matters, suppose that M is of the form L_α , as constructed in V , where α is a countable ordinal in V . The existence of models of ZFC of this form follows, for instance, from the existence of large cardinals (see section 6 below), so we certainly cannot rule out this possibility. Since we want to build a model $M[c]$ of ZFC that contains a new infinite binary sequence c and all the elements of M , it will have to contain $L_\alpha(c)$, i.e., all sets that can be constructed in fewer than α steps starting with c . But if c is a sequence that encodes α , as above, then $M[c]$ cannot equal $L_\alpha(c)$ and still be a model of ZFC, since this would imply that $L_\alpha(c)$ contained itself. If we try to circumvent the problem by adding more sets to $M[c]$ so that it becomes a model of ZFC, then we may end up with $M[c] = L_\gamma$ for some ordinal γ greater than α . And this is not good for

our purposes since CH holds in all models of ZFC of the form L_γ . The conclusion is that we cannot just pick an *arbitrary* c that is not in M : we will have to choose it very carefully.

The key idea is that c should be “generic,” meaning that it should have no special property that singles it out. The reason for this is that if, as before, $M = L_\alpha$, and we want to ensure that $M[c] = L_\alpha(c)$ is still a model of ZFC, then we do not want c to have any special property that would interfere in the construction of $M[c]$ and cause some ZFC axiom not to hold any more. To accomplish this we build c little by little so that it avoids all the special properties that could possibly have any undesirable effect on $M[c]$. For example, if we do not want c to encode the ordinal α in the manner sketched above, we simply set some $c(n)$ equal to 0 for some n such that $g(n) \in A$.

Of course, if we have built up the first N binary digits of c and φ is a property that holds for all real numbers that begin with those N digits, then we cannot avoid φ without undoing our previous work. Let us call a property *avoidable* if every finite binary sequence p can be extended to a finite binary sequence q such that no infinite sequence that extends q has the property. For instance, the property “all terms in the sequence are zero” is avoidable, while the property “there are ten consecutive ones in the sequence” is not avoidable.

A real number c is called *generic*, or *Cohen*, over M if it avoids all avoidable properties that can be defined in M , that is, properties that can be defined by means of formulas that may mention sets in M . It is easy to see that c cannot belong to M , since if it did then the property “is equal to c ” would be definable in M , and it is certainly avoidable.

Why should a generic real number exist? Once again, we use the fact that M is countable. From this it follows that there are only countably many avoidable properties. If we enumerate them as $\varphi_1, \varphi_2, \dots$, then we can pick a finite sequence q_1 such that no infinite extension of q_1 satisfies φ_1 . Then we can extend q_1 to q_2 such that no infinite extension of q_2 satisfies φ_2 . Continuing in this way we create an infinite binary sequence c that does not have any of the properties φ_i . In other words, it is generic.

Now let $M[c]$ be the set of all sets that can be constructed, using c and the elements of M as parameters, in as many steps as the ordinals of M . For instance, if M were of the form L_α , then $M[c]$ would just be $L_\alpha(c)$. The model $M[c]$ is called a *Cohen-generic extension* of M .

It turns out that, miraculously, $M[c]$ is a model of ZFC. Moreover, it has the same ordinals as M and, therefore, it is not of the form L_γ , for any ordinal γ . In particular, when we build L inside $M[c]$, c does not belong to it. These statements are by no means easy to prove, but very roughly what Cohen showed was that a formula φ is true in $M[c]$ if and only if there is an initial segment p of c that “forces” φ to be true. Moreover, the relation “ p forces φ to be true,” which relates finite binary sequences to formulas and is written $p \Vdash \varphi$, can be defined in M . Therefore, to know whether a statement φ is true in $M[c]$ one just needs to check whether there is an initial segment p of c such that $p \Vdash \varphi$. In particular, using this result one can prove that $M[c]$ satisfies the ZFC axioms.

In order to build a model where CH fails, one adds not just one generic real number but \aleph_2^M of them, where \aleph_2^M is the ordinal that plays the role of \aleph_2 in M . That is, it is the second uncountable cardinal in M . This need not be the real \aleph_2 , and indeed it will not be if, for instance, M is of the form L_α for some countable ordinal α in V . Adding \aleph_2^M generic real numbers can be done by finitely approximating any finite number of them while avoiding all avoidable properties they could have. Thus, instead of finite binary sequences we now work with finite sets of finite binary sequences indexed by ordinals less than \aleph_2^M . A generic object will be a sequence $\langle c_\alpha : \alpha < \aleph_2^M \rangle$ of Cohen reals over M , all different, and so CH is false in the generic extension $M[\langle c_\alpha : \alpha < \aleph_2^M \rangle]$.

However, there is an important point that needs to be addressed. When we add the new real numbers to M , it is important that the \aleph_2 of the new expanded model is the same as \aleph_2^M . Otherwise, CH might hold in the expanded model and our work would have been wasted. Fortunately, this is true, but again we must use the facts about forcing to prove it.

The same kind of forcing argument allows one to construct models where the cardinality of \mathbb{R} is \aleph_3 , or \aleph_{27} , or any other cardinal of uncountable *cofinality*, i.e., any uncountable cardinal that is not the least upper bound of countably many smaller cardinals. The cardinality of the continuum is, therefore, undetermined by ZFC. Furthermore, since CH holds in Gödel’s constructible universe L and fails in the model constructed by Cohen using forcing, it is undecidable in ZFC.

Cohen also used forcing to prove that AC is independent of ZF. Since AC holds in L , this amounted to constructing a model of ZF in which AC was false. He did this by adding a countable collection $\langle c_n : n \in \mathbb{N} \rangle$ of

generic real numbers to a countable model M of ZF. To see why this works, let N be the smallest submodel of $M[\langle c_n : n \in \mathbb{N} \rangle]$ that contains all the ordinals and the unordered set $A = \{c_n : n \in \mathbb{N}\}$. Thus, N is just $L(A)$, as built inside $M[\langle c_n : n \in \mathbb{N} \rangle]$. One can then show that N is a model of ZF, but that in N there is no well-ordering of A . The reason is that any well-ordering of A would be definable in $L(A)$ with a finite number of ordinals and finitely many elements of A as parameters, and then each one of the c_n would in its turn be definable by indicating its ordinal position in the well-ordering. But since the whole sequence of c_n s is generic over L , no formula can distinguish one of the c_n s from another unless they appear as parameters in the formula. Since we can choose two different c_n s that do not appear as parameters in the definition of the well-ordering of A , and that well-ordering distinguishes all the c_n s from each other, we have a contradiction. Therefore, the set A cannot be well-ordered, so AC does not hold.

Immediately after Cohen's proof of the independence of AC from ZF and of CH from ZFC, a result for which he got the Fields Medal in 1966, many set theorists started developing the forcing technique in its full generality (notably Azriel Lévy, Dana Scott, Joseph Shoenfield, and Robert Solovay) and began to apply it to other well-known mathematical problems. For instance, Solovay constructed a model of ZF in which every set of real numbers is LEBESGUE MEASURABLE [III.57], thereby showing that AC is necessary for the existence of non-measurable sets. He also constructed a model of ZFC where every *definable* set of real numbers is Lebesgue measurable; therefore, nonmeasurable sets, although they can be proved to exist (see the example in section 6.1 below), cannot be explicitly given; Solovay and Stanley Tennenbaum developed the theory of iterated forcing and used it to prove the consistency of Suslin's hypothesis (see section 10 below); Adrian Mathias proved the consistency of the infinitary form of RAMSEY'S THEOREM [IV.19 §2.2]; Saharon Shelah proved the undecidability of the Whitehead problem in group theory; and Richard Laver proved the consistency of the Borel conjecture; to cite just a few remarkable examples from the 1970s.

The forcing technique now pervades all of set theory. It continues to be a research area of great interest, very sophisticated from the technical point of view and of great beauty. It keeps producing important results, with applications in many areas of mathematics, such as topology, combinatorics, and analysis. Especially influential has been the development over the last

twenty-five years of the theory of *proper forcing*, introduced by Shelah. Proper forcing has proved very useful in the context of forcing iterations, and in the formulation and study of new *forcing axioms*, which will be dealt with in section 10, as well as in the analysis of *cardinal invariants* of the continuum. These are uncountable cardinals associated with various topological or combinatorial properties of the real line that can consistently take different values in different models obtained by forcing. An example of a cardinal invariant is the least number of null sets needed to cover the real line. Another important development has been the use of *class forcing* by Anthony Dodd and Ronald Jensen for coding the universe into a single real number, which shows that, amazingly, one can always use forcing to turn any model M into a model of the form $L(r)$ for some real number r . A more recent contribution is the invention by W. Hugh Woodin of new powerful forcing notions associated with the theory of large cardinals (see the next section), which have provided new insights into the continuum hypothesis (see the end of section 10).

The large number of independence results obtained by forcing have made very clear that the axioms of ZFC are insufficient to answer many fundamental mathematical questions. Thus, it is desirable to find new axioms that, once added to ZFC, will provide a solution to some of those questions. We shall discuss some candidates in the next few sections.

PUP: Tim thinks that it's the 'results' that have made things clear, rather than their sheer number, so 'have' OK?

6 Large Cardinals

As we have already seen, the collection of all ordinal numbers cannot form a set. But if it did, then to that set there would correspond an ordinal number κ . This ordinal would coincide with the κ th cardinal \aleph_κ , since otherwise \aleph_κ would be a larger ordinal. Moreover, V_κ would be a model of ZFC. We cannot prove in ZFC that there is an ordinal κ with these properties, for then we would have proved in ZFC that ZFC has a model, which is impossible by Gödel's second incompleteness theorem. So, why do we not add to ZFC the axiom that says that there is a cardinal κ such that V_κ is a model of ZFC?

This axiom, with the further requirement that κ be *regular*, that is, not the limit of fewer than κ smaller cardinals, was proposed in 1930 by SIERPIŃSKI [VI.77] and TARSKI [VI.87], and it is the first of the *large-cardinal axioms*. A cardinal κ with those properties is called *inaccessible*.

Other notions of large cardinals, which implied inaccessibility, kept appearing during the twentieth century. Some of them originated in generalizations to uncountable sets of the infinite version of Ramsey's theorem, which states that if each (unordered) pair of elements of ω (i.e., of natural numbers) is painted either red or blue, then there is an infinite subset X of ω such that all pairs of elements of X have the same color. The natural generalization of the theorem to ω_1 turns out to be false. However, on the positive side, Paul Erdős and Richard Rado proved that for every cardinal $\kappa > 2^{\aleph_0}$, if each pair of elements of κ is painted either red or blue, then there is a subset X of κ of size ω_1 such that all pairs of elements of X have the same color. This is one of the landmark results of the *partition calculus*, an important area of combinatorial set theory developed mainly by the Hungarian school, led by Erdős and András Hajnal. The problem of whether Ramsey's theorem can be generalized to some uncountable cardinal leads naturally to cardinals that are called *weakly compact*. A cardinal κ is weakly compact if it is uncountable and satisfies the strongest possible Ramsey-type theorem: whenever all pairs of elements of κ are painted either red or blue, there is a subset X of κ of size κ such that all pairs of elements of X have the same color. Weakly compact cardinals are inaccessible, so their existence cannot be proved in ZFC. Moreover, it turns out that below the first weakly compact cardinal, assuming it exists, there are many inaccessible cardinals, so the existence of a weakly compact cardinal cannot be proved even if one assumes the existence of inaccessible cardinals.

The most important large cardinals, the *measurable cardinals*, are much larger than the weakly compact ones, and were discovered in 1930 by Stanisław Ulam.

6.1 Measurable Cardinals

A set A of real numbers is a **BOREL SET** [III.57] if it can be obtained in countably many steps starting from the open intervals and applying the two operations of taking complements and countable unions. It is *null*, or has *measure zero*, if for every $\varepsilon > 0$ there is a sequence of open intervals I_0, I_1, I_2, \dots such that $A \subseteq \bigcup_n I_n$ and $\sum_n |I_n| < \varepsilon$. It is *Lebesgue measurable* if it is almost a Borel set, that is, if it differs from a Borel set by a null set. To each measurable set A corresponds a number $\mu(A) \in [0, \infty]$, its *measure*, that is invariant under translation of A and is *countably additive*, that is, the measure of a countable union of measurable pairwise-disjoint sets is the sum of their measures. Moreover,

the measure of an interval is its length (see MEASURES [III.57]).

One can prove in ZFC that there exist non-Lebesgue-measurable sets of real numbers. For example, the following set was discovered in 1905 by Giuseppe Vitali. Define two elements of the closed interval $[0, 1]$ to be equivalent if they differ by a rational, and let A be a subset of $[0, 1]$ that contains precisely one element from each equivalence class. This requires one to make a large number of choices, which can be done by AC. To see that A is not measurable, consider for each rational p the set $A_p = \{x + p : x \in A\}$. Any two of these sets are disjoint, because of the way we built A . Let B be the union of all A_p over all rational numbers p in the interval $[-1, 1]$. A cannot have measure zero, for then B itself would have measure zero, and this is impossible because $[0, 1] \subseteq B$. On the other hand, A cannot have positive measure either, since then B would have infinite measure, and this is impossible because $B \subseteq [-1, 2]$.

Since measurable sets are closed under taking complements and countable unions, all Borel sets are measurable. In 1905 LEBESGUE [VI.72] showed that there are measurable sets that are not Borel. While reading Lebesgue's work, Mikhail Suslin noticed that Lebesgue had made a mistake in claiming that continuous images of Borel sets are Borel. Indeed, Suslin soon found a counterexample, which led eventually to the discovery of a new natural hierarchy of sets of reals beyond the Borel sets, the so-called *projective sets*. These are the sets that can be obtained from the Borel sets by taking continuous images and complements (see section 9 below). In 1917 Nikolai Luzin showed that all continuous images of Borel sets, the *analytic sets*, are also measurable. If a set is measurable, then so is its complement, so all complements of analytic sets, the *coanalytic sets*, are also Lebesgue measurable. It is therefore natural to ask whether we can continue like this. In particular, are continuous images of coanalytic sets, or Σ_2^1 sets, as they are known, also measurable? The answer to this question turns out to be undecidable in ZFC: in L there are Σ_2^1 sets that are not Lebesgue measurable, and with forcing one can construct models where all Σ_2^1 sets are measurable.

The proof given above of the existence of a non-Lebesgue-measurable set of reals hinges on the fact that Lebesgue measure is translation invariant. In fact, the proof shows that there cannot be any countably additive translation-invariant measure that extends Lebesgue measure and measures all sets of reals. Thus,

a natural question, known as *the measure problem*, is whether, if one drops the requirement of translation invariance, there can exist some countably additive measure that extends Lebesgue measure and measures all sets of reals. If such a measure exists, then the cardinality of the continuum cannot be \aleph_1 , nor \aleph_2 , nor any \aleph_n with $n < \omega$, etc. In fact, Ulam proved in 1930 that a positive solution to the measure problem implies that the cardinality of \mathbb{R} is extremely large: it is greater than or equal to the least uncountable regular cardinal that is a limit of smaller cardinals. He also proved that the existence of a nontrivial countably additive measure on *any* set implies either a positive solution to the measure problem, or that there exists an uncountable cardinal κ with a (nontrivial) $\{0, 1\}$ -valued κ -additive measure that measures all its subsets. Such a cardinal is called *measurable*. If κ is measurable, then it is weakly compact, and therefore inaccessible. In fact, the set of weakly compact cardinals smaller than κ has measure 1, and so κ is itself the κ th weakly compact cardinal. It follows that the existence of a measurable cardinal cannot be proved in ZFC, even if one adds the axiom that inaccessible, or weakly compact, cardinals exist (unless, of course, ZFC plus the existence of such cardinals is inconsistent). A complete clarification of the measure problem was finally provided by Solovay, who showed that if the solution is positive, then there is an inner model with a measurable cardinal. Conversely, if there is a measurable cardinal, then one can build a forcing extension where the measure problem has a positive solution.

An unexpected consequence of the existence of a measurable cardinal is that the universe V cannot be L : that is, there are nonconstructible sets, and even nonconstructible real numbers. In fact, if there is a measurable cardinal, then V is much larger than L . For instance, the first uncountable cardinal, \aleph_1 , is an inaccessible cardinal in L .

After the invention of forcing and the subsequent avalanche of independence results, the hope arose that axioms asserting the existence of large cardinals, like measurable cardinals, would settle some of the questions that, thanks to the forcing technique, had been proved undecidable in ZFC. It was soon shown, however, by Lévy and Solovay, that large-cardinal axioms could not settle CH, as one could easily use forcing to change the cardinality of the continuum and make CH hold or fail without destroying the large cardinals. But Solovay proved in 1969 that, surprisingly, if there exists a measurable cardinal, then all Σ_2^1 sets of real num-

bers are Lebesgue measurable. So, while the axiom that asserts the existence of a measurable cardinal cannot settle the size of the continuum, it has a profound effect on its structure. It is indeed astonishing that measurable cardinals, so far away from the sets of real numbers in the universe V , have such a strong influence on their basic properties. While the relationship between large cardinals and the structure of the continuum is not yet fully understood, great progress has been made in the last thirty years through the work done in *descriptive set theory* and *determinacy*, which will be described in sections 8 and 9 below.

Some of the deepest and most technically difficult work in set theory is currently devoted to the construction and analysis of canonical inner models for large cardinals. These are analogues of L for large cardinals, that is, they are models built in some canonical way that contain all the ordinals and are transitive (i.e., they contain all elements of their elements), and in which certain large cardinals exist. The larger the cardinal, the more difficult it is to build the model. This work is known as the *inner model program*.

One of the striking consequences of the inner model program is that it provides a way of measuring the *consistency strength* of virtually any set-theoretic statement φ , using large cardinals. That is, there are large-cardinal axioms A_1 and A_2 such that the consistency of ZFC plus φ implies that of ZFC plus A_1 and is implied by the consistency of ZFC plus A_2 . We refer to A_1 as a *lower bound* for the consistency of φ and to A_2 as an *upper bound*. In the fortunate cases when the lower and upper bounds coincide, we obtain an exact measure of the consistency strength of φ . An upper bound A_2 is usually obtained by forcing over a model of ZFC plus A_2 , whereas a lower bound A_1 is obtained by inner model theory. Earlier in this section we saw that the consistency strength of a positive solution to the measure problem is exactly that of the existence of a measurable cardinal. We shall see another important example in the next section.

Knowing upper and lower bounds for the consistency strength of set-theoretic statements—or, even better, knowing their exact consistency strength—is extremely useful for comparing them. Indeed, if the lower bound for a sentence φ is greater than the upper bound for another sentence ψ , then we can conclude, via Gödel's incompleteness theorem, that ψ does not imply φ .

7 Cardinal Arithmetic

Beyond the continuum hypothesis, understanding the behavior of the exponential function 2^κ for arbitrary infinite cardinals κ has been a motivating force in set theory. Cantor proved that $2^\kappa > \kappa$ for all κ , and Dénes König proved that the cofinality of 2^κ is always greater than κ : that is, 2^κ is not the limit of fewer than κ smaller cardinals. The GCH, which, as we saw, holds in the constructible universe L , states precisely that 2^κ has the least possible value, namely, the least cardinal greater than κ , usually denoted by κ^+ . One might think that, as in the case of 2^{\aleph_0} , by forcing it should be possible to build models of ZFC where 2^κ takes any prescribed value, subject only to the necessary requirement that its cofinality should be greater than κ . This is true for cardinals κ that are *regular*, that is, not the limit of fewer than κ smaller cardinals. Indeed, William Easton showed that for any function F on the regular cardinals such that $\kappa \leq \lambda$ implies $F(\kappa) \leq F(\lambda)$ and $F(\kappa)$ has cofinality greater than κ , there is a forcing extension of L in which $2^\kappa = F(\kappa)$, for all regular κ . So, for instance, one can build a model of ZFC where $2^{\aleph_0} = \aleph_7$, $2^{\aleph_1} = \aleph_{20}$, $2^{\aleph_2} = \aleph_{20}$, $2^{\aleph_3} = \aleph_{101}$, etc. This shows that the behavior of the exponential function for infinite regular cardinals is totally undetermined in ZFC, and anything possible can be attained by forcing.

But how about nonregular cardinals? Nonregular cardinals are called *singular*. Thus, an infinite cardinal κ is *singular* if it is the supremum of fewer than κ smaller cardinals. For instance, \aleph_ω , being the supremum of the \aleph_n , $n \in \mathbb{N}$, is the first singular cardinal. Determining the possible values of the exponential function at singular cardinals is a very hard problem that has generated much important research and involves, quite surprisingly, the necessary use of large cardinals.

Using a *supercompact cardinal*, which is a measurable cardinal with certain further properties that make it much larger than ordinary measurable cardinals, Matthew Foreman and Woodin built a model of ZFC in which GCH fails everywhere, i.e., $2^\kappa > \kappa^+$ for all cardinals κ . But curiously, the value of the exponential function at a singular cardinal of *uncountable* cofinality is somehow determined by its values at smaller regular cardinals. Indeed, in 1975, Jack Silver proved that if κ is a singular cardinal of uncountable cofinality and $2^\alpha = \alpha^+$ for all $\alpha < \kappa$, then $2^\kappa = \kappa^+$. That is, if the GCH holds below κ , then it also holds at κ . That this is also the case for singular cardinals of *countable* cofinality is a consequence of the *singular cardinal hypoth-*

I	n_0	n_2	n_4	\dots	n_{2k}	\dots
II	n_1	n_3	n_5	\dots	n_{2k+1}	\dots

Figure 3 A run of the infinite game associated with a set $A \subseteq [0, 1]$.

esis (SCH), a general principle weaker than the GCH that completely determines singular cardinal exponentiation, relative to exponentiation for regular cardinals. A special case of SCH is the following. *If $2^{\aleph_n} < \aleph_\omega$ for all finite n , then $2^{\aleph_\omega} = \aleph_{\omega+1}$.* So, in particular, if the GCH holds below \aleph_ω , then it must hold at \aleph_ω . Shelah used his powerful “PCF theory” to obtain the unexpected result that if $2^{\aleph_n} < \aleph_\omega$ for all n , then $2^{\aleph_\omega} < \aleph_{\omega_4}$. So, if GCH holds below \aleph_ω , then there is a bound (in ZFC!) on the possible values of 2^{\aleph_ω} . But can this value actually be greater than the least possible one, namely $\aleph_{\omega+1}$? In particular, can the GCH first fail at \aleph_ω ? The answer is yes, but large cardinals are needed. Indeed, on the one hand Menachem Magidor proved the consistency of the first failure of GCH at \aleph_ω , assuming the consistency of the existence of a supercompact cardinal. Thus, the existence of a supercompact cardinal is an *upper bound* for the failure of SCH. On the other hand, using inner model theory, Dodd and Jensen showed that large cardinals are required for this to happen. An exact measure of the consistency strength of the failure of SCH was later established by Moti Gitik.

8 Determinacy

It turns out that the existence of very large cardinals, such as supercompact cardinals, has a dramatic effect on the properties of sets of real numbers, especially when they can be defined in some simple way. The link between the two appears through the analysis of certain infinite two-player games that are associated with sets of real numbers. Given a subset A of $[0, 1]$, consider the following infinite game associated with A : there are two players, I and II, who alternately choose a number n_i that equals either 0 or 1. To begin with, player I plays n_0 , then player II plays n_1 , to which I answers by playing n_2 , and so on. A run of the game is displayed in figure 3. At the end of the run, the players have produced an infinite binary sequence: n_0, n_1, n_2, \dots . This sequence can be regarded as the binary expansion of a real number r in $[0, 1]$. Player I wins the game if r belongs to A and player II wins otherwise.

For example, if A is the interval $[0, \frac{1}{2}]$, then a winning strategy for player I is simply to start by play-

PUP: Tim confirms that ‘ \aleph_{20} ’ being on the right-hand side of two different equalities is correct.

PUP: Tim would prefer not to spell this out but thinks that putting it into inverted commas makes it clear you don’t need to know what the letters stand for. OK?

ing 0, whereas if $A = [0, \frac{1}{4})$, then player II wins the game by playing 1 in her first move. But for most games, the question of who wins is not decided after any finite number of moves. For instance, if A is the set of rational points of $[0, 1]$, then one can easily see that player II has a strategy for winning the game (for example, whatever player I does, player II will win if she plays 01001000100001...), but she will not win at any finite stage of the run.

The game is *determined* if one of the two players has a winning strategy. Formally, a *strategy* for player II is a function f that assigns 0 or 1 to each finite binary sequence of odd length. It is a *winning strategy* if player II always wins the game if she plays $f(n_0, n_1, \dots, n_{2k})$ in her k th turn, whatever moves are made by player I. Similarly, one can define a winning strategy for I. We say that the set A is *determined* if the game associated with A is determined. One might guess that every game is determined, but actually it is quite easy, using AC, to prove the existence of a game that is not determined.

It turns out that the determinacy of the games associated with certain classes of sets of reals implies that all sets in the class have properties similar to those of the Borel sets. For example, the *axiom of determinacy* (AD), which asserts that all sets of reals are determined, implies that every set of reals is Lebesgue measurable, has the property of Baire (i.e., differs from an open set by a set of first category), and has the perfect set property (i.e., contains a perfect set if it is uncountable). To give the flavor of a typical argument, let us indicate why every set A of reals is Lebesgue measurable.

First, one observes that it is enough to show that if all measurable subsets of A are null, then A itself must be null. And for this one plays, for every $\varepsilon > 0$, the *covering game* for A and ε . In this game, player I plays so that the sequence $a = \langle n_0, n_2, n_4, \dots \rangle$ represents an element of A , and player II plays (binary encodings of) finite unions of rational intervals, with measures adding up to at most ε , while attempting to cover a . It can be shown that if every measurable subset of A is null, then player I cannot have a winning strategy. So by AD there must be a winning strategy for II. Using this strategy one can show that the outer measure of A is at most ε . And since this works for all $\varepsilon > 0$, A must be null.

While AD rules out the existence of badly behaved sets of reals, it implies the negation of AC, so AD is inconsistent with ZFC. However, weaker versions of AD are compatible with, and even follow from, ZFC. Indeed,

Donald Martin proved in 1975 that ZFC implies that every Borel set is determined. Moreover, if there exists a measurable cardinal, then every analytic set, and therefore also every coanalytic set, is determined. A natural question, therefore, is whether the existence of larger cardinals implies the determinacy of more complex sets such as the Σ_2^1 sets.

The intimate connection between large cardinals and the determinacy of simple sets of reals was first made explicit by Leo Harrington, who showed that the determinacy of all analytic sets is in fact equivalent to a large-cardinal principle slightly weaker than the existence of a measurable cardinal. As we shall shortly see, large cardinals imply the determinacy of certain simply definable sets of reals, the so-called projective sets, while the determinacy of those sets implies in turn the existence of the same kind of large cardinals in some inner models.

9 Projective Sets and Descriptive Set Theory

As we have seen, very basic questions about sets of real numbers can be extremely hard to answer. However, it often turns out to be possible to answer them for sets that occur “in nature,” or that can be explicitly described. This raises the hope that one might be able to prove facts about definable sets of reals that cannot be proved for arbitrary sets.

The study of the structure of definable sets of reals is the subject of *descriptive set theory*. Examples of such sets are the Borel sets, and also the *projective* sets, which are sets that can be obtained from Borel sets by taking continuous images and complements. An equivalent definition of the projective sets is that they are subsets of \mathbb{R} that can be obtained from closed subsets of \mathbb{R}^n by a mixture of projecting to a lower dimension and taking complements. To see how this relates to definability, consider projecting a subset $A \subset \mathbb{R}^2$ down to the x -axis. The result will be the set of all x such that there exists y with $(x, y) \in A$. Thus, projection corresponds to existential quantification. Taking complements corresponds to negation, so one can combine the two and obtain universal quantification as well. One can therefore think of a projective set as a set that is definable from a closed set.

Since analytic sets are continuous images of Borel sets, they are projective. And so are the complements of the analytic sets, the coanalytic sets, and the continuous images of coanalytic sets, the Σ_2^1 sets. More

complex projective sets are obtained by taking complements of Σ_2^1 sets, the so-called Π_2^1 sets, their continuous images, called Σ_3^1 , etc. The projective sets form a hierarchy of increasing complexity, in accordance with the number of steps (always finite) that are necessary to obtain them from the Borel sets. Many sets of reals that appear naturally in usual mathematical practice are projective. Moreover, the results and techniques of descriptive set theory, although originally developed for the study of sets of reals, also apply to definable sets in any *Polish space* (a separable and complete-metrizable space). These include basic examples such as \mathbb{R}^n , \mathbb{C} , separable BANACH SPACES [III.64], etc., where projective sets arise in a very natural way. For example, in the space $C[0, 1]$ of continuous real-valued functions on $[0, 1]$ with the sup norm, the set of everywhere differentiable functions is coanalytic, and the set of functions that satisfy the mean-value theorem is Π_2^1 . Thus, since descriptive set theory deals with rather natural sets in Polish spaces of general mathematical interest, it is not surprising that it has found many applications in other areas of mathematics such as harmonic analysis, group actions, ergodic theory, and dynamical systems.

Classical results of descriptive set theory are that all analytic sets, and hence also all coanalytic sets, are Lebesgue measurable and have the Baire property, and that all uncountable analytic sets contain a perfect set. However, as we have already pointed out, one cannot prove in ZFC that all Σ_2^1 sets have those properties, since in L there are counterexamples. By contrast, if there exists a measurable cardinal, then they do have them. But what about more complex projective sets?

The theory of projective sets is closely tied to large cardinals. On the one hand, Solovay showed that if the existence of an inaccessible cardinal is consistent, then so is the statement that every projective set of reals is Lebesgue measurable, has the Baire property, etc. On the other hand, Shelah showed, quite unexpectedly, that the inaccessible cardinal is necessary, in the sense that if all Σ_3^1 sets are Lebesgue measurable, then \aleph_1 is an inaccessible cardinal in L .

Nearly all the classical properties of Borel and analytic sets are shared by the projective sets, assuming that they are determined. So since the determinacy of all projective sets cannot be proved in ZFC and since it allows for the extension of the theory of Borel and analytic sets to all projective sets in a very elegant and satisfactory way, it constitutes an excellent candidate for a

new set-theoretic axiom. This axiom is known as *projective determinacy* (PD). It implies, for instance, that every projective set is Lebesgue measurable, has the Baire property, and has the perfect set property. In particular, since every uncountable perfect set has the same cardinality as \mathbb{R} , it implies that there is no projective counterexample to CH.

One of the most remarkable advances in set theory over the last twenty years is the proof that PD follows from the existence of large cardinals. Martin and John Steel proved in 1988 that if there exist infinitely many so-called *Woodin cardinals*, then PD holds. Woodin cardinals lie between measurable and supercompact in the hierarchy of large cardinals. Subsequently, Woodin showed that, surprisingly, the hypothesis that for each n it is consistent that there exist n Woodin cardinals is necessary in order to obtain the consistency of PD. Thus the existence of infinitely many Woodin cardinals is a sufficient, and essentially necessary, assumption for extending the classical theory of Borel and analytic sets to all projective sets of reals, and more generally to all projective sets in Polish spaces.

In spite of the enormous success of the known large-cardinal axioms, not only in descriptive set theory but also in many other areas of mathematics, their status as true axioms of set theory is still a matter of debate. This is more so in the case of very large cardinals such as the supercompact ones, the reason being that there is as yet no inner model theory available for them, which means that there is not even strong evidence for their consistency. However, it should be noted that, as Harvey Friedman has shown, large cardinals are necessary even for proving quite simple-looking and rather natural statements about finite functions on the integers, which provides evidence for their essential role in even the most basic parts of mathematics. Another shortcoming of the known large-cardinal axioms is that they cannot decide some fundamental questions. The most conspicuous is CH, but there are others.

10 Forcing Axioms

Another old and basic question about the continuum that the known large-cardinal axioms cannot solve is *Suslin's hypothesis* (SH). Cantor had proved that every linearly ordered set that is dense (i.e., any two distinct elements have another element in between), complete (i.e., every nonempty subset with an upper bound has a supremum), separable (i.e., contains a dense countable subset), and without endpoints is order-isomorphic to

PUP: Tim confirms that this jargon is necessary here.

the real line. In 1920 Suslin conjectured that if instead of separability one assumes the weaker *countable chain condition*, or CCC, which demands that every pairwise-disjoint collection of open intervals should be at most countable, then it must still be isomorphic to \mathbb{R} . The importance of SH for the development of set theory is that it led to the discovery of a new class of axioms, the so-called *forcing axioms*.

In 1967, Solovay and Tennenbaum used forcing to construct a model in which SH holds. The idea is to use the forcing to destroy any counterexamples that there might be to SH. But when one does this one may create new ones, and the result is that one needs to force again and again, transfinitely many times. The iteration of forcing is technically cumbersome and difficult to control, for many unwanted things can happen at the limit stages. For instance, ω_1 may be “collapsed,” i.e., it may become countable.

Fortunately, these difficulties can be dealt with. In general, a forcing argument involves a partially ordered set. (In the case we looked at earlier, it was the set of all finite binary sequences, with $p < q$ if p was a proper initial segment of q .) If one starts with a model where GCH holds, uses only partial orderings that are CCC—that is, in which every set of incompatible elements is countable—and takes so-called *direct limits* at the limit stages, then in ω_2 steps one can destroy all counterexamples so that SH holds in the final model. On the other hand, Jensen proved in 1968 that a counterexample to SH exists in L , thereby proving the undecidability of SH in ZFC.

From the construction of Solovay and Tennenbaum, Martin isolated a new principle now known as *Martin’s axiom* (MA), which generalizes the well-known *Baire category theorem*. The latter states that in every compact Hausdorff topological space, the intersection of a countable collection of dense open sets is nonempty. MA says the following:

In every compact Hausdorff CCC topological space, the intersection of \aleph_1 dense open sets is nonempty.

The condition that the space be CCC (i.e., every collection of pairwise-disjoint open sets is countable) is necessary, for without it the statement is false. It is easy to see that MA implies the negation of CH, for if there are only \aleph_1 real numbers, then the intersection of the \aleph_1 dense open sets $\mathbb{R} \setminus \{r\}$, as r ranges over all the real numbers, is empty. However, MA does not decide the cardinality of \mathbb{R} .

MA has been used with great success to solve many questions that are undecidable in ZFC. For example, it implies SH and that every Σ_2^1 set is Lebesgue measurable. But is MA really an axiom? In what sense, if any, is it a natural, or at least plausible, assumption about sets? Is the fact that it decides many ZFC undecidable questions sufficient for it to be accepted as being on a par with the ZFC axioms or the axioms of large cardinals? We shall come back to this.

MA has many different equivalent formulations. The original formulation of Martin was more closely connected with forcing—hence the term *forcing axiom*. Roughly speaking it said that if you have a CCC partial order, then you can avoid \aleph_1 avoidable properties, and not just countably many. This allows one to prove the existence of generic subsets of the partial order, over models M of size \aleph_1 .

Stronger forcing axioms can be obtained by expanding the class of partial orderings to which MA applies while keeping the axiom consistent. An important such strengthening is the *proper forcing axiom* (PFA), which is formulated for partial orderings that are *proper*. Properness is a property weaker than the CCC that was discovered by Shelah and is particularly useful when working with complicated forcing iterations. The strongest possible forcing axiom of this type was discovered by Foreman, Magidor, and Shelah in 1988. It is called *Martin’s maximum* (MM) and is consistent with ZFC, assuming the consistency of a supercompact cardinal.

Both MM and PFA have striking consequences. For example, PFA, and therefore also MM, implies the axiom of projective determinacy (PD), the singular cardinal hypothesis (SCH), and that the cardinality of \mathbb{R} is \aleph_2 .

An advantage of forcing axioms is that one can apply them without having to go into the details of forcing, just as \diamond and \square save one from having to go into the details of constructible sets. A very good example of this is PFA and some combinatorial principles derived from it, like the so-called *open coloring axiom*, which have been used with great success by Stevo Todorcevic to solve many outstanding problems in general topology and infinite combinatorics.

As we have already pointed out, forcing axioms are not as intuitively evident as the ZFC axioms, or even the axioms of large cardinals, so one can ask to what extent they should be considered as true axioms of set theory rather than just useful principles for showing that certain statements are consistent with ZFC. In the case of MA and some weaker forms of PFA and MM,

PUP: Tim thinks
this is OK as it is.

some justification for their being taken as true axioms is based on the fact that they are equivalent to principles of *generic absoluteness*. That is, they assert, under certain restrictions that are necessary to avoid inconsistency, that *everything that might exist, does exist*. More precisely, if some set having certain properties could be forced to exist over V , then a set having the same properties already exists (in V). So, like the axioms of large cardinals, they are maximality principles, i.e., they attempt to make V as large as possible.

For example, MA is equivalent to the assertion that if a set X having some properties that depend exclusively on subsets of ω_1 could be forced to exist over V using a CCC partial ordering \mathbb{P} , then such an X already exists in V . This characterization of MA in terms of generic absoluteness provides some justification for regarding MA as a true axiom of set theory. The analogous principle of generic absoluteness, but for proper partial orderings instead of CCC, is known as the *bounded proper forcing axiom* (BPFA). Although weaker than PFA, BPFA is strong enough to decide many questions that the large-cardinal axioms are unable to settle. Most notably, Justin Moore has recently proved, following a series of results by Woodin, David Asperó, and Todorcevic, that BPFA implies that the cardinality of \mathbb{R} is \aleph_2 .

To finish, let us briefly mention some deep results that establish strong underlying connections between large cardinals, inner models, determinacy, forcing axioms, generic absoluteness, and the continuum. These results hold under the assumption that for every ordinal α there exists a Woodin cardinal greater than α .

The first one, due to Shelah and Woodin, is that the theory of $L(\mathbb{R})$ is generically absolute. That is, all sentences with real numbers as parameters that would hold in the $L(\mathbb{R})$ of *any* generic extension of V are already true in the real $L(\mathbb{R})$. This kind of generic absoluteness implies that all sets of reals in $L(\mathbb{R})$, and in particular the projective sets, are Lebesgue measurable, have the Baire property, etc. Furthermore, by refining the Martin–Steel result that large cardinals imply PD, Woodin showed that in $L(\mathbb{R})$ every set of reals is determined.

Another result of Woodin is that there is an axiom, which he calls $(*)$, that is intended to play the role for subsets of ω_1 that PD plays for sets of natural numbers, in the sense that it decides “practically all” questions about those sets. Of course, no consistent axiom can really decide *all* questions that refer only to subsets of ω_1 , since by Gödel’s incompleteness theorem there will

always be undecidable arithmetical statements. So, to formulate precisely the notion of *deciding practically all questions*, Woodin introduces a new logic, called Ω -logic, that strengthens ordinary first-order logic. One of the main features of Ω -logic is that the valid statements in Ω -logic are generically absolute. Under suitable large-cardinal hypotheses, $(*)$ is consistent in Ω -logic and decides in Ω -logic all questions that refer only to subsets of ω_1 . The main open problem is the Ω -conjecture, whose formulation is quite technical and beyond the scope of this article. If the Ω -conjecture is true, then *any* axiom compatible with the existence of large cardinals that decides all questions that depend exclusively on subsets of ω_1 in Ω -logic must imply the negation of CH. Thus, the theories ZFC plus CH and ZFC plus not-CH are not equally reasonable from the point of view of Ω -logic, since in the presence of large cardinals CH puts an unnecessary limitation on the possibility of settling all natural questions about subsets of ω_1 .

11 Final Remarks

In this short account of set theory, we have reviewed some of the key developments since its beginnings in the late nineteenth century. What started in the hands of Cantor as a mathematical theory of transfinite numbers has developed to become a general theory of infinite sets and a foundation for mathematics. The fact that it has been possible to unify all of classical mathematics into one single theoretical framework, the ZFC axiom system, is certainly remarkable. But beyond this, and most importantly, the techniques developed by set theory, such as constructibility, forcing, infinite combinatorics, the theory of large cardinals, determinacy, the descriptive theory of definable sets in Polish spaces, etc., have turned it into a discipline of great depth and beauty, with fascinating results that stimulate and challenge our imagination, and with numerous applications in areas such as algebra, topology, real and complex analysis, functional analysis, and measure theory. In the twenty-first century, the ideas and techniques generated within set theory will surely continue to contribute to the solution of outstanding mathematical problems, old as well as new, and will help mathematicians gain an ever deeper insight into the complexities and vastness of the mathematical universe.

PUP: Tim has deleted some of this sentence but is resistant to the specific changes suggested by the proofreader. OK as it is now?

Further Reading

- Foreman, M., and A. Kanamori, eds. 2008. *Handbook of Set Theory*. New York: Springer.
- Friedman, S. D. 2000. *Fine Structure and Class Forcing*. De Gruyter Series in Logic and Its Applications, volume 3. Berlin: Walter de Gruyter.
- Hrbacek, K., and T. Jech. 1999. *Introduction to Set Theory*, 3rd edn. (revised and expanded). New York: Marcel Dekker.
- Jech, T. 2003. *Set Theory*, 3rd edn. New York: Springer.
- Kanamori, A. 2003. *The Higher Infinite*, 2nd edn. Springer Monographs in Mathematics. New York: Springer.
- Kechris, A. S. 1995. *Classical Descriptive Set Theory*. Graduate Texts in Mathematics. New York: Springer.
- Kunen, K. 1980. *Set Theory: An Introduction to Independence Proofs*. Amsterdam: North-Holland.
- Shelah, S. 1998. *Proper and Improper Forcing*, 2nd edn. New York: Springer.
- Woodin, W. H. 1999. *The Axiom of Determinacy, Forcing Axioms, and the Nonstationary Ideal*. De Gruyter Series in Logic and Its Applications, volume 1. Berlin: Walter de Gruyter.
- Zeman, M. 2001. *Inner Models and Large Cardinals*. De Gruyter Series in Logic and Its Applications, volume 5. Berlin: Walter de Gruyter.

IV.23 Logic and Model Theory

David Marker

1 Languages and Theories

Mathematical logic is the study of formal languages that are used to describe mathematical structures and what these can tell us about the structures themselves. We can learn a lot about a formal language by investigating which of its sentences are true for the structure it describes, and we can learn a lot about the structure by investigating the subsets of it that can be defined using the language. In this article, we shall see several examples of languages and the structures that they are used to describe. We shall also see instances of the remarkable phenomenon that theorems in logic can sometimes be used to prove “purely mathematical” results that seem to have nothing to do with logic. This introductory section briefly introduces some of the basic ideas that will be needed to understand the later sections.

All the formal languages that we consider will be extensions of a basic logical language that we shall denote by \mathcal{L}_0 . The statements, or *formulas*, of this language are made up of the following components: *variables*, which are denoted by letters of the alphabet such as

x or y , or letters with subscripts such as v_1, v_2, \dots ; the *parentheses* “(” and “)”; the *equality symbol* “=”; the *logical connectives* $\wedge, \vee, \neg, \rightarrow, \leftrightarrow$, which we read as “and,” “or,” “not,” “implies,” and “if and only if”; and the *quantifiers* \exists and \forall , which we read as “there exists” and “for all.” (If these symbols are unfamiliar to you, then you should read THE LANGUAGE AND GRAMMAR OF MATHEMATICS [I.2] before attempting to read this article.) Here are a couple of formulas of \mathcal{L}_0 :

- (i) $\forall x \forall y \exists z (z \neq x \wedge z \neq y)$;
- (ii) $\forall x (x = y \vee x = z)$.

The first of these says that if any object exists at all then there are at least three objects, and the second says that y and z are the only objects. There is an important difference between the two formulas: the variables x, y , and z that occur in the first formula are all *bound* variables, which means that they are all attached to quantifiers, whereas in the second formula, only the variable x is bound, while the variables y and z are *free*. This means that the first formula expresses a statement about some mathematical structure, while the second is a statement about not just a structure but also the particular elements y and z .

There are various rules that allow one to build larger formulas out of smaller ones. We will not give them all, but for example if ϕ and ψ are formulas, then $\neg\phi$, $\phi \vee \psi$, $\phi \wedge \psi$, $\phi \rightarrow \psi$, and $\phi \leftrightarrow \psi$ are all formulas. In general, if ϕ is built out of smaller formulas ϕ_1, \dots, ϕ_n using logical connectives (and parentheses), then we call ϕ a *Boolean combination* of ϕ_1, \dots, ϕ_n . Another important way to modify a formula is quantification: if $\phi(x)$ is a formula involving a free variable x , then $\forall x\phi(x)$ and $\exists x\phi(x)$ are both formulas.

The formulas just discussed are “purely logical,” which makes them not very useful for describing interesting mathematical structures. Suppose, for example, that we wanted to study real solutions to algebraic and exponential equations over the FIELD [I.3 §2.2] of real numbers. We can think of this as studying the “mathematical structure”

$$\mathbb{R}_{\text{exp}} = (\mathbb{R}, +, \cdot, \text{exp}, <, 0, 1),$$

where the right-hand side is a septuple that consists of the set \mathbb{R} of real numbers, the binary operations of addition and multiplication, the EXPONENTIAL FUNCTION [III.25], the “less than” relation, and the real numbers 0 and 1.

The various components of this structure are of course related to each other in many ways, but we can-

not express these relationships unless we are prepared to extend the basic language \mathcal{L}_0 . For example, if we wanted to write, in a formal way, the statement that the exponential function turns addition into multiplication, then the obvious thing to write down would be

$$(i) \quad \forall x \forall y \quad \exp(x) \cdot \exp(y) = \exp(x + y).$$

Here we have two quantifiers, two bound variables x and y , and the equals sign, but the rest of the formula involves extraneous elements such as “+”, “·”, and “exp”. Thus, to discuss the structure \mathbb{R}_{\exp} , we extend the language \mathcal{L}_0 to a language \mathcal{L}_{\exp} , by adding in the symbols “+”, “·”, “exp”, “<”, “0”, and “1”. Of course, these come with various syntactic rules that reflect the fact that “+” is a binary operation, “exp” is a function, and so on. For instance, these rules would allow us to write $\exp(x + y) = z$ but would forbid us to write $\exp(x = y) + z$.

Here are three more \mathcal{L}_{\exp} -formulas:

- (ii) $\forall x (x > 0 \rightarrow \exists y \exp(y) = x)$;
- (iii) $\exists x x^2 = -1$;
- (iv) $\exists y y^2 = x$.

We interpret these formulas as the assertions “for all positive x , there is a y such that $e^y = x$,” “ -1 is a square,” and “ x is a square.” The first three formulas above are declarative statements about the structure \mathbb{R}_{\exp} . Formulas (i) and (ii) are true in \mathbb{R}_{\exp} , while (iii) is false. Formula (iv) is different because x is a free variable: thus, it expresses a property of x . (For instance, it is true if $x = 8$, but false if $x = -7$.) A *sentence* is defined to be a formula with no free variables. If ϕ is an \mathcal{L}_{\exp} -sentence, then ϕ is either true or false in \mathbb{R}_{\exp} .

If ϕ is a formula with free variables x_1, \dots, x_n , and a_1, \dots, a_n are real numbers, then we write $\mathbb{R}_{\exp} \models \phi(a_1, \dots, a_n)$ if the formula ϕ is true for the particular sequence (a_1, \dots, a_n) . We think of the formula as defining the set

$$\{(a_1, \dots, a_n) \in \mathbb{R}^n : \mathbb{R}_{\exp} \models \phi(a_1, \dots, a_n)\},$$

that is, the set of all sequences (a_1, \dots, a_n) for which the formula is true when you set x_i to equal a_i for every i . For example, the formula

$$\exists z (x = z^2 + 1 \wedge y = z \cdot \exp(\exp(z)))$$

defines the parametrized curve

$$\{(t^2 + 1, te^{e^t}) : t \in \mathbb{R}\}.$$

For another example, one that illustrates an important point, let us consider the structure $(\mathbb{Z}, +, \cdot, 0, 1)$: that is, the integers, with addition, multiplication, 0, and 1. The language used to describe this structure is the *language of rings*, $\mathcal{L}_{\text{rng}} = \mathcal{L}(+, \cdot, 0, 1)$. (The notation here lists the symbols that we add to the basic language \mathcal{L}_0 .) The language \mathcal{L}_{rng} has no symbol for the usual ordering on \mathbb{Z} , but, surprisingly, this ordering can nevertheless be defined in terms of \mathcal{L}_{rng} . (To appreciate the nonobviousness of this fact, the reader is encouraged to try to work out why it is true before reading on.)

The trick is to use a well-known theorem due to LAGRANGE [VI.22], which asserts that every nonnegative integer is a sum of four squares. It follows that the statement $x \geq 0$ can be defined by the formula

$$\exists y_1 \exists y_2 \exists y_3 \exists y_4 \quad x = y_1^2 + y_2^2 + y_3^2 + y_4^2.$$

(Of course, we are also using the fact that a negative integer cannot be written as a sum of four squares. Note too that a similar trick would work even if all one knew was that every nonnegative integer was a sum of a hundred squares.) Once one has a way of expressing the statement that x is nonnegative, it is easy to define the symbol “<”. The interesting aspect of this is that the reformulation was not obvious—it depended on a genuine mathematical theorem.

It is important to understand that formulas are restricted in several ways, of which two stand out in particular.

- Formulas are finite. We do not allow formulas like $\forall x > 0 (x < 1 \vee x < 1 + 1 \vee x < 1 + 1 + 1 \vee \dots)$, which would express the fact that \mathbb{R} has the so-called Archimedean property. (If we did, then it would be much easier to define “<” above.)
- Quantifiers range over *elements* of the structure, and not subsets. This rules out a “second-order” formula such as

$$\forall S \subseteq \mathbb{R} \quad (\text{if } S \text{ is bounded above, then } S \text{ has a least upper bound}),$$

which would express the completeness of \mathbb{R} by quantifying over all subsets S of \mathbb{R} . Since we look just at “first-order” formulas, what we are studying is often called *first-order logic*.

Now that we have seen some examples of languages, let us discuss them more generally. A *language* is basically something like \mathcal{L}_{\exp} or \mathcal{L}_{rng} above: that is, a set

of symbols (combined with the basic logical symbols) together with some rules concerning their use. If \mathcal{L} is a language, then an \mathcal{L} -structure is a mathematical structure in which all the sentences of \mathcal{L} can be interpreted. (This concept will become clearer in a moment, when we give a couple of examples.) An \mathcal{L} -theory T is just a set of \mathcal{L} -sentences, which one can think of as axioms that an \mathcal{L} -structure might or might not satisfy. A *model* of T is then an \mathcal{L} -structure \mathcal{M} in which all the sentences of T , suitably interpreted, are true. For instance, the structure was a model for the formulas (i) and (ii) of the language \mathcal{L}_{exp} that we discussed earlier. (Another model for the same two formulas would be one in which we replaced the exponential function by the function 2^x and interpreted “exp” as referring to that function instead.)

The justification for the word “theory” is clearer in another example, the language of GROUPS [I.3 §2.1], $\mathcal{L}_{\text{grp}} = \mathcal{L}(\circ, e)$. Here, \circ is a binary operation symbol and e is a constant. We might look at the theory T_{grp} consisting of the sentences

- (i) $\forall x \forall y \forall z \ x \circ (y \circ z) = (x \circ y) \circ z$;
- (ii) $\forall x \ x \circ e = e \circ x = x$;
- (iii) $\forall x \exists y \ x \circ y = y \circ x = e$;

which are the usual axioms for groups.

In order to interpret this language in some mathematical structure \mathcal{M} we need \mathcal{M} to consist of a set M , a binary operation $f : M^2 \rightarrow M$, and an element $a \in M$. We then interpret “ \circ ” as referring to f , “ e ” as referring to the element a , and quantification as being over the set M . Thus, for example, the interpretation of (iii) is that for every x in M there exists a y in M such that $f(x, y) = a$. Under this interpretation of the symbols of \mathcal{L}_{grp} , the structure \mathcal{M} becomes an \mathcal{L}_{grp} -structure. This \mathcal{L}_{grp} -structure is a model of T_{grp} if in addition the sentences (i), (ii), and (iii) are all true. Since sentences (i)–(iii) are the axioms for groups, a model of T_{grp} is nothing other than a group.

We say that an \mathcal{L} -sentence ϕ is a *logical consequence* of a theory T , and write $T \models \phi$, if ϕ is true in every model of T . That is, $T \models \phi$ if ϕ is true in every structure in which all the sentences of T are true. Thus, the symbol “ \models ” has two different meanings, according to whether there is a structure or a theory on the left-hand side. However, these two meanings are closely related in that they are both concerned with truth in models: $\mathcal{M} \models \phi$ means that ϕ is true in the model \mathcal{M} , and $T \models \phi$, as we have just said, means that ϕ is true in every pos-

sible model of T . Either way, the symbol “ \models ” stands for a “semantic” notion of entailment.

Returning to the example of groups, if ϕ is a sentence in \mathcal{L}_{grp} , then $T_{\text{grp}} \models \phi$ if and only if ϕ is true for every group. So, for instance,

$$T_{\text{grp}} \models \forall x \forall y \forall z (xy \neq xz \vee y = z),$$

because if x , y , and z are elements of any group and $xy = xz$, then we can multiply both sides on the left by the inverse of x to deduce that $y = z$.

We can now describe some of the basic problems in logic.

- (i) Given an \mathcal{L} -theory T , can we decide if a sentence ϕ is a logical consequence of T , and if so how?
- (ii) Given an interesting mathematical structure, like \mathbb{R}_{exp} , or $(\mathbb{N}, +, \cdot, 0, 1)$, or the complex field, and a language \mathcal{L} that describes the structure, can we determine which \mathcal{L} -sentences are true of the structure?
- (iii) Given a structure described by a language, do the subsets of the structure that can be defined in the language have special properties? Are they in some sense “simple”? For example, earlier we saw how to use \mathcal{L}_{exp} to define a certain curve in the plane. Now consider a very complicated set such as a CANTOR SET [III.17] or the MANDELBROT SET [IV.14 §2.8]. Is it possible to prove that these sets *cannot* be defined in \mathcal{L}_{exp} because they are “too complex” in some sense?

2 Completeness and Incompleteness

Let T be an \mathcal{L} -theory and let ϕ be an \mathcal{L} -sentence. To show that $T \models \phi$, we must show that ϕ holds in every model of T . Checking all models of T sounds like a daunting task, but fortunately it is not necessary, since instead we can use a *proof*. One of the first tasks in mathematical logic is to say precisely what this means.

Suppose, then, that \mathcal{L} is some language and that T is a set of sentences in \mathcal{L} , i.e., an \mathcal{L} -theory. Suppose also that ϕ is a formula of \mathcal{L} . Informally speaking, a proof of ϕ assumes the statements of T and ends up establishing ϕ . We express this idea formally as follows. A *proof of ϕ from T* is a finite sequence of \mathcal{L} -formulas ψ_1, \dots, ψ_m (which one can think of as the lines of the proof) with the following properties:

- (i) each ψ_i is either a logical axiom, or a sentence of T , or a formula that follows from the previous formulas $\psi_1, \dots, \psi_{i-1}$ by means of simple logical rules;

(ii) $\psi_m = \phi$.

We shall not say precisely what a “simple logical rule” is, but three examples are

- from ϕ and ψ it follows that $\phi \wedge \psi$;
- from $\phi \wedge \psi$ it follows that ϕ ;
- from $\phi(x)$ it follows that $\exists v \phi(v)$.

The other possible rules are similarly elementary.

There are three points about proofs that need to be stressed. The first is that they are finite, which may seem too obvious to mention but is important because it has a number of consequences that are not obvious. The second is that proof systems have to be *sound*: if there is a proof of ϕ from T , then ϕ is true in every model of T . To put this more succinctly, let us introduce the notation $T \vdash \phi$ for the statement that there is a proof of ϕ from T . Then soundness is the assertion that if $T \vdash \phi$ then $T \models \phi$. This is why we can prove that ϕ is true in every model of T by finding a proof rather than by looking at all the models. The third point is that it is easy to check whether a sequence of sentences is a proof. More precisely, there is an algorithm that can look at a sequence ψ_1, \dots, ψ_m and decide whether it really is a proof of ϕ from T .

It is not too surprising that if ϕ can be proved from T , then ϕ is true in all models of T . Much more remarkable is that the converse is also true: if ϕ cannot be proved from T , then there must be a model of T in which ϕ is false. This tells us that two very different notions—the finitistic, syntactic notion of “proof” and the semantic notion of “logical consequence,” which concerns truth in models—always agree. This result is known as Gödel’s completeness theorem. Here is its formal statement.

Theorem. *Let T be an \mathcal{L} -theory and let ϕ be an \mathcal{L} -sentence. Then $T \models \phi$ if and only if $T \vdash \phi$.*

Suppose that T is a simple theory like T_{grp} , where there is an algorithm to decide whether a sentence is in T . (In the case of T_{grp} this algorithm is particularly simple, but some theories might have infinitely many sentences.) We could write a computer program which, given a formula ϕ as its input, would systematically generate all possible proofs σ from T and check to see whether σ was a proof of ϕ . If such a program finds a proof of ϕ , then it halts and tells us that $T \models \phi$. We say that $\{\phi : T \models \phi\}$ is *recursively enumerable*.

However, one might hope for more. If $T \not\models \phi$, our program above will go on searching forever, so it will

never tell us that there is no proof of ϕ . We say that an \mathcal{L} -theory T is *decidable* if there is a computer program which, when given an \mathcal{L} -sentence ϕ as input, will always halt and tell us, one way or another, whether $T \models \phi$. Such a program would have to be cleverer than the one that just checks all possible proofs σ , and unfortunately such a program does not have to exist: as GÖDEL [VI.92] proved in his famous INCOMPLETENESS THEOREM [V.18], many important theories are undecidable. Here is a first version of his theorem, concerning the *theory of the natural numbers* (or theory of \mathbb{N} for short), which means the set of all sentences in the language \mathcal{L}_{rng} that are true of the structure $(\mathbb{N}, +, \cdot, 0, 1)$.

Theorem. *The theory of the natural numbers is undecidable.*

At first, this might seem rather strange: after all, if T is the theory of \mathbb{N} , then T contains all true sentences about \mathbb{N} . So a sentence ϕ is provable from T if and only if it has a one-line proof (the line being ϕ itself). However, this does not make ϕ decidable, because the theory T is very complicated and there is no algorithm for deciding whether ϕ belongs to T .

One approach to proving the incompleteness theorem is to associate a natural number with each computer program in such a way that statements about programs can be recast as statements about natural numbers. The theory of \mathbb{N} then determines whether a program P halts on input x , thus solving what is known as the *halting problem*. Since the halting problem was shown by TURING [VI.94] to be undecidable (a sketch of the proof can be found in THE INSOLUBILITY OF THE HALTING PROBLEM [V.23]), it follows that the theory of \mathbb{N} is undecidable.

How can we understand the theory of \mathbb{N} ? One might hope to find a much smaller theory that yielded the same true sentences. That is, we could try to find a simple set of axioms about \mathbb{N} that we know are true and hope that every true sentence follows from these axioms. A good candidate is *first-order Peano arithmetic*, or PA. This is a theory in the language $\mathcal{L}(+, \cdot, 0, 1)$ that involves a few simple axioms about addition and multiplication, such as

$$\forall x \forall y \ x \cdot (y + 1) = x \cdot y + x,$$

together with axioms for induction.

Why do we need more than one axiom of induction? The reason is that the obvious statement that expresses the principle of mathematical induction, namely

$$\forall A \ (0 \in A \wedge \forall x \ x \in A \rightarrow x + 1 \in A) \rightarrow \forall x \ x \in A,$$

is not a first-order sentence, because the quantifier is applied to all subsets A of \mathbb{N} . (It is also not a sentence in \mathcal{L}_{rng} since it uses the symbol “ \in ”, but this is a less fundamental problem.) To get around this difficulty, one has a separate axiom of induction for each formula ϕ . It is the assertion that

$$[\phi(0) \wedge \forall x (\phi(x) \rightarrow \phi(x+1))] \rightarrow \forall x \phi(x).$$

In words, this says that if $\phi(0)$ is true and $\phi(x+1)$ is true whenever $\phi(x)$ is true, then $\phi(x)$ is true for every x in \mathbb{N} .

Most of number theory can be formalized in PA and one might hope that $\text{PA} \vdash \phi$ for every ϕ that is true in \mathbb{N} . Sadly, this is not true. Here is a second version of Gödel’s incompleteness theorem. Recall that the notation $\mathbb{N} \models \psi$ means simply that ψ is true in \mathbb{N} .

Theorem. *There is a sentence ψ such that $\mathbb{N} \models \psi$ but $\text{PA} \not\vdash \psi$.*

Another way to state this result is to say that there is a sentence ψ such that $\text{PA} \not\vdash \psi$ and $\text{PA} \not\vdash \neg\psi$. To see that this is an equivalent statement, let ψ be any sentence. Then precisely one of ψ and $\neg\psi$ is true. Therefore, if the theorem is false, then PA must prove either ψ or $\neg\psi$. But this means that we can decide which by simply going through all possible proofs in PA until we find a proof of ψ or a proof of $\neg\psi$.

Gödel’s original example of a true but unprovable sentence was a self-referential sentence that effectively asserted

“I am not provable from PA.”

More precisely, he found a sentence ψ for which he was able to show that ψ is true in \mathbb{N} if and only if ψ is not provable from PA. With more work he showed that there is a sentence that asserts

“PA is consistent”

that is unprovable from PA. The somewhat artificial and metamathematical nature of these sentences might lead one to hope that all “mathematically interesting” sentences about \mathbb{N} are settled by PA. However, more recent work has shown that even this is a forlorn hope, since there are undecidable statements related to RAMSEY’S THEOREM [IV.19 §2.2] in finite combinatorics.

Undecidability also appears in number theory in a very basic way. *Hilbert’s tenth problem* asked if there is an algorithm to decide whether a polynomial $p(X_1, \dots, X_n)$ with integer coefficients has an integer zero. Davis, Matijasevic, Putnam, and Robinson showed that the answer is no.

Theorem. *For any recursively enumerable $S \subseteq \mathbb{N}$ there is $n > 0$ and $p(X, Y_1, \dots, Y_n) \in \mathbb{Z}[X, Y_1, \dots, Y_n]$ such that $m \in S$ if and only if $p(m, Y_1, \dots, Y_n)$ has an integer zero.*

Since the halting problem provides an undecidable recursively enumerable set, the answer to Hilbert’s tenth problem is no. An important open question is whether there is an algorithm to decide if a polynomial with *rational* coefficients has a *rational* zero. Hilbert’s tenth problem is also discussed in THE INSOLUBILITY OF THE HALTING PROBLEM [V.23], and other interesting examples of undecidability can be found in GEOMETRIC AND COMBINATORIAL GROUP THEORY [IV.10].

3 Compactness

A theory T is called *satisfiable* if there are structures that satisfy all of the sentences in T (that is, if T has a model), and we call T *consistent* if we cannot derive a contradiction from T . Since our proof system is sound, any satisfiable theory is consistent. On the other hand if T is not satisfiable, then every sentence ϕ is a logical consequence of T , for the trivial reason that there are no models of T in which ϕ is required to be true. But the completeness theorem then tells us that $T \vdash \phi$ for every ϕ . Choosing ϕ to be some contradictory statement, of the form $\psi \wedge \neg\psi$, for instance, we see that T is inconsistent. This way of reformulating the completeness theorem has the following simple consequence, called the *compactness theorem*, which turns out to be surprisingly important, as we shall see.

Theorem. *If every finite subset of T is satisfiable, then T is satisfiable.*

The reason this is true is that if T is not satisfiable then it is inconsistent (as we have just seen), which means that a contradiction can be proved from T . Since this proof, like all proofs, must be finite, it involves only finitely many sentences from T . Therefore, T has a finite subset that implies a contradiction, which contradicts our assumption that all finite subsets of T are satisfiable.

Although the compactness theorem is an easy consequence of the completeness theorem, it has many immediate intriguing consequences and lies at the heart of many constructions in model theory. Here are two simple applications that show that theories have many models that you might not expect. If \mathcal{M} is some \mathcal{L} -structure, let us write $\text{Th}(\mathcal{M})$ for *the theory of \mathcal{M}* :

that is, for the set of all \mathcal{L} -sentences that are true in \mathcal{M} . We also extend our earlier notation $\mathcal{M} \models \phi$ from single formulas to collections of formulas, so if \mathcal{M} is an \mathcal{L} -structure and T is an \mathcal{L} -theory, then $\mathcal{M} \models T$ means that every sentence of T is true in \mathcal{M} , or in other words that \mathcal{M} is a model of T .

Corollary. *There exists an \mathcal{L}_{exp} -structure \mathcal{M} containing an infinite element a (which means that $a > 1$, $a > 1 + 1$, $a > 1 + 1 + 1$, etc.), such that $\mathcal{M} \models \text{Th}(\mathbb{R}_{\text{exp}})$.*

That is, there is a structure \mathcal{M} in which all the true statements about the structure \mathbb{R}_{exp} are still true, but \mathcal{M} is different from \mathbb{R}_{exp} because it contains an infinite element. To prove this, we add one more constant symbol c to our language and consider the theory T that consists of all the statements of $\text{Th}(\mathbb{R}_{\text{exp}})$ (that is, all true statements about \mathbb{R}_{exp}), together with the infinite sequence of statements $c > 1$, $c > 1 + 1$, $c > 1 + 1 + 1$, and so on. If Δ is any finite subset of T , then we can make \mathbb{R} a model of Δ simply by interpreting c as a sufficiently large real number—large enough to satisfy all the statements of the form $c > 1 + 1 + \dots + 1$ that belong to Δ . Since we can model every finite subset Δ of T , the compactness theorem tells us that we can model T itself. If $\mathcal{M} \models T$, then the element named by c must be infinite.

The element $1/a$ will be an *infinitesimal* element of \mathcal{M} (which means that it satisfies statements that effectively say that it is smaller than $1/n$ for every positive integer n). This observation is the first step toward a rigorous development of calculus with infinitesimals.

For another example, let $\mathcal{L}_{\text{rng}} = \mathcal{L}(+, \cdot, 0, 1)$ be the language of rings. Let T be the set of \mathcal{L} -sentences that are true in every finite field. We call T the *theory of finite fields*. Recall that a field is said to have *characteristic* p if p is the smallest positive integer (which has to be prime) such that $1 + 1 + \dots + 1 = 0$ in the field, where the number of 1s in the sum is p . If there is no such p , then the field is said to have *characteristic zero*. Thus, the fields \mathbb{Q} , \mathbb{R} , and \mathbb{C} all have characteristic zero.

Corollary. *There is a field F with characteristic zero such that $F \models T$.*

This result tells us that there is no possible set of axioms that characterizes the finite fields: given any set of statements that are true in all finite fields, there is an infinite field in which they are also all true. To prove it, we look at the theory T' that consists of T together with the statements $1 + 1 \neq 0$, $1 + 1 + 1 \neq 0$, and so on. Any finite set of statements in T' will be true of a finite

field of sufficiently large characteristic, and thus satisfiable. By the compactness theorem T' is satisfiable, but a model of T clearly has to have characteristic zero.

The compactness theorem can sometimes be used to show the existence of interesting algebraic bounds. The next result allows us to deduce from HILBERT'S NULLSTELLENSATZ [V.20] a stronger “quantitative version.” It is our first example of a statement that does not appear to be logical in nature but which can be proved using logic. Recall that a field is *algebraically closed* if every polynomial with coefficients in the field has a root in the field. (THE FUNDAMENTAL THEOREM OF ALGEBRA [V.15] is the assertion that \mathbb{C} is an algebraically closed field.)

Proposition. *For any three positive integers n , m , d there is a positive integer l such that if K is an algebraically closed field and f_1, \dots, f_m are polynomials in n variables with coefficients in K , degree at most d and no common zero, then there are polynomials g_1, \dots, g_m of degree at most l such that $\sum g_i f_i = 1$.*

Hilbert's Nullstellensatz itself is the same statement but without the extra information about the degrees of the polynomials g_i .

To see how the proposition is proved, we will restrict our attention to the case $n = d = 2$. This is just for notational simplicity: the proof is almost identical in larger cases. For each i between 1 and m let

$$F_i = a_i X^2 + b_i Y^2 + c_i XY + d_i X + e_i Y + f_i.$$

For each k write down a formula ϕ_k that asserts that there are no polynomials G_1, \dots, G_m with degree at most k such that $1 = \sum F_i G_i$. Let T be the theory of algebraically closed fields with the formulas ϕ_1, ϕ_2, \dots and the assertion that the polynomials F_1, \dots, F_m have no common zero. If there is no positive integer l satisfying the conclusion of the proposition, then every finite subset of T is satisfiable. Hence, by the compactness theorem, T is satisfiable. If $K \models T$, then F_1, \dots, F_m are polynomials over an algebraically closed field with no common zero, but it is impossible to find polynomials G_1, \dots, G_m such that $\sum G_i F_i = 1$. This contradicts Hilbert's Nullstellensatz.

Notice that in the above argument we did not say anything about the dependence of l on n , m , and d . This is because the proof does not actually find a bound: it merely shows that some sort of bound must exist. However, good explicit bounds were recently discovered—see ALGEBRAIC GEOMETRY [IV.4] for more details.

4 The Complex Field

A surprising counterpoint to Gödel's incompleteness theorem is a result of TARSKI [VI.87], which states that the theories of the fields of real and complex numbers *are* decidable. The key to these results is a method known as *quantifier elimination*. If we have a formula without quantifiers that concerns the natural numbers, then it is easy to decide whether it is true or false. The negative solution to Hilbert's tenth problem shows that as soon as we start adding existential quantifiers (as we do if, for example, we assert that a polynomial has a zero), then we leave the realm of decidability.

Thus, if we want to show that a formula is decidable, it will be very useful if we can find an equivalent formula that does not have quantifiers. And in some settings, this turns out to be possible. For example, let $\phi(a, b, c)$ be the formula

$$\exists x \, ax^2 + bx + c = 0.$$

The usual rule for solving quadratics tells us that, as long as $a \neq 0$, this is true in \mathbb{R} if and only if $b^2 \geq 4ac$. Therefore, $\mathbb{R} \models \phi(a, b, c)$ if and only if

$$[(a \neq 0 \wedge b^2 - 4ac \geq 0) \vee (a = 0 \wedge (b \neq 0 \vee c = 0))].$$

As for the complex numbers, it is easy to see that $\mathbb{C} \models \phi(a, b, c)$ if and only if

$$a \neq 0 \vee b \neq 0 \vee c = 0.$$

In either case, ϕ is equivalent to a formula with no quantifiers.

For a second example, let $\phi(a, b, c, d)$ be the formula

$$\begin{aligned} \exists x \exists y \exists u \exists v \, (xa + yc = 1 \wedge xb + yd = 0 \\ \wedge ua + vc = 0 \wedge ub + vd = 1). \end{aligned}$$

The formula $\phi(a, b, c, d)$ is the obvious way of asserting that the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is invertible. However, by the DETERMINANT [III.15] test, we know that, for any field F , $F \models \phi(a, b, c, d)$ if and only if $ad - bc \neq 0$. Thus the existence of an inverse can be expressed by the quantifier-free formula $ad - bc \neq 0$.

Tarski proved that we can *always* eliminate quantifiers in algebraically closed fields.

Theorem. *For any \mathcal{L}_{rng} -formula ϕ there is a quantifier-free formula ψ such that ϕ is equivalent to ψ in every algebraically closed field.*

Furthermore, Tarski gave an explicit algorithm for eliminating the quantifiers.

The equivalent quantifier-free formulas above were both finite Boolean combinations of formulas of the

form $p(v_1, \dots, v_n) = q(v_1, \dots, v_n)$, where p and q are polynomials in n variables with integer coefficients. It is not hard to see that this is true of any quantifier-free \mathcal{L}_{rng} -formula. It follows that a quantifier-free \mathcal{L}_{rng} -sentence is particularly simple: if no free variables are allowed and no quantifiers are allowed, then there cannot be any variables! Therefore, the polynomials p and q have to be constant, which means that a quantifier-free \mathcal{L}_{rng} -sentence is a finite Boolean combination of formulas of the form $k = l$ (where this should be regarded as an abbreviation for $1 + 1 + \dots + 1 = 1 + 1 + \dots + 1$, with k 1s on the left-hand side and l 1s on the right-hand side).

This leads to the decidability result. If we want to know whether $\mathbb{C} \models \phi$, then we use Tarski's algorithm to convert ϕ into an equivalent quantifier-free sentence. But the very simple form of such sentences makes their truth or falsity easy to decide.

In the remainder of this section, we shall discuss a number of other consequences of Tarski's theorem. The first is that sentences in the language \mathcal{L}_{rng} cannot distinguish between different algebraically closed fields of the same characteristic. That is, if ϕ is any \mathcal{L}_{rng} -sentence that is true for some algebraically closed field of characteristic p (where p is allowed to be zero), then it is true in every algebraically closed field of characteristic p .

To see why this is true, let K and F be two algebraically closed fields of characteristic p , and suppose that $K \models \phi$ (or in other words that ϕ is true of K). Let k be the field \mathbb{Q} if the characteristic is zero and the field with p elements otherwise. Tarski's theorem tells us that there is a quantifier-free sentence ψ that is equivalent to ϕ in all algebraically closed fields of characteristic p . However, the extremely simple nature of the quantifier-free sentences of \mathcal{L}_{rng} means that their truth or falsity in any given field depends only on the elements 0, 1, $1 + 1$, and so on. Therefore,

$$K \models \psi \Leftrightarrow k \models \psi \Leftrightarrow F \models \psi.$$

Since $K \models \phi$ and ϕ and ψ are equivalent in all algebraically closed fields of characteristic p , it follows that $F \models \phi$ as well.

A consequence of this theorem is that an \mathcal{L}_{rng} -sentence ϕ is true of the complex numbers if and only if it is true of the algebraic numbers \mathbb{Q}^{alg} . (Recall that these are all roots of polynomials with integer coefficients. As one would expect, the algebraic numbers form an algebraically closed field, though this is not a wholly obvious fact.) Thus, rather surprisingly, if we

wish to prove something about \mathbb{Q}^{alg} , we have the option of working in \mathbb{C} and using the methods of complex analysis; similarly, if we want to prove something about \mathbb{C} we can, if it makes things easier, work in \mathbb{Q}^{alg} and use number-theoretic methods.

Combining these ideas with the completeness theorem gives another useful tool. If ϕ is any \mathcal{L}_{rng} -sentence, then the following are equivalent:

- (i) ϕ is true in every algebraically closed field of characteristic zero;
- (ii) for some $m > 0$, ϕ is true in every algebraically closed field of characteristic $p > m$;
- (iii) there are arbitrarily large p such that ϕ is true in some algebraically closed field of characteristic p .

Let us see why this is so. Suppose first that ϕ is true in every algebraically closed field of characteristic 0. The completeness theorem then implies that there is a *proof* of ϕ from the axioms for algebraically closed fields combined with the sentences $1 \neq 0$, $1 + 1 \neq 0$, $1 + 1 + 1 \neq 0$, and so on. Since proofs are finite sequences of formulas, there must be some m such that the proof used only the first m of these sentences (not necessarily all of them). If p is some prime bigger than m , then this proof shows that ϕ holds in algebraically closed fields of characteristic p , since all the sentences we used are true in such fields.

We have just shown that (i) implies (ii). It is obvious that (ii) implies (iii). To see that (iii) implies (i), let us suppose that (i) fails, so that there is an algebraically closed field of characteristic zero in which $\neg\phi$ is true. Then, by the principle we proved earlier, $\neg\phi$ is true in every algebraically closed field of characteristic zero. Thus, since (i) implies (ii), there is an m such that $\neg\phi$ is true in every algebraically closed field of characteristic $p > m$. Therefore (iii) fails.

An interesting application of this theorem was found by Ax. It is another example of a statement that has nothing to do with logic, but which can be proved using logical tools. It is perhaps more striking than the previous example because in this case one does not even feel with hindsight that the statement did after all have some logical content.

Theorem. *If a polynomial map from \mathbb{C}^n to \mathbb{C}^n is an injection, then it must also be a surjection.*

The basic thought behind the proof of this result is very simple indeed: what is remarkable is that it is of any help. It is the observation that if k is a finite field,

then every injective polynomial map from k^n to k^n is a surjection. This is true because every injection from a finite set to itself is automatically a surjection.

How do we exploit this observation? Well, the previous results tell us that, in several situations, statements are true for one field if and only if they are true for another. We shall use these results to transfer our problem from \mathbb{C} , where it is hard, to a finite field k , where it is trivial. The first step is a routine exercise: one shows that for each positive integer d there is a sentence ϕ_d in \mathcal{L}_{rng} that expresses the fact that every injective polynomial map from F^n to F^n , with the n polynomials all of degree at most d , is surjective. We would like to prove that all the sentences ϕ_d are true when $F = \mathbb{C}$.

The equivalences in the previous theorem imply that it is enough to prove that the sentences ϕ_d are true when F is the field $\mathbb{F}_p^{\text{alg}}$, the algebraic closure of the p -element field. (It can be shown that any field F is contained in an algebraically closed field. Roughly speaking, the *algebraic closure* of F is the smallest algebraically closed field that contains F .) Suppose, then, that some ϕ_d fails for $\mathbb{F}_p^{\text{alg}}$. Then there must be an injective polynomial map f from $(\mathbb{F}_p^{\text{alg}})^n$ to $(\mathbb{F}_p^{\text{alg}})^n$ that is not surjective. Since every finite subset of $\mathbb{F}_p^{\text{alg}}$ is contained in a finite subfield, there is a finite subfield k such that all the n polynomials used to define f have coefficients in k , from which it follows that f maps k^n to k^n . Moreover, by enlarging k if necessary, we can ensure that there is an element of k^n that is not in the image of f . But now we have succeeded in transferring ourselves to a finite field: this function $f : k^n \rightarrow k^n$ is an injection between finite sets that is not a surjection, which is a contradiction.

Quantifier elimination has other useful applications. Let F be a field, let K be a subfield of F , let $\psi(v_1, \dots, v_n)$ be a quantifier-free formula, and let a_1, \dots, a_n be elements of K . Since, as we have already mentioned, quantifier-free formulas are just Boolean combinations of equalities between polynomials, the statement $\psi(a_1, \dots, a_n)$ involves just the elements of K , and is therefore true in K if and only if it is true in F . By quantifier elimination, if K and F are algebraically closed, then the same is true for *all* formulas ψ , and not just those that are quantifier free. From this observation we can prove the “weak version” of Hilbert’s Nullstellensatz. (For the proof, we shall need to assume a certain degree of familiarity with the basics of RING THEORY [III.83]. We shall also write $K[X]$ for the polynomial ring $K[X_1, \dots, X_n]$ and \bar{v} for the n -tuple (v_1, \dots, v_n) .)

Proposition. Suppose that K is an algebraically closed field, P is a prime ideal in $K[X]$, and g is a polynomial in $K[X]$ that does not belong to P . Then there is some $a = (a_1, \dots, a_n)$ in K^n such that $f(a) = 0$ for every f that belongs to P , and such that $g(a) \neq 0$.

Proof. Let F be the algebraic closure of the fraction field of the integral domain $K[X]/P$. We can view F as an extension field of K with a natural homomorphism $\eta : K[X] \rightarrow F$. Let $b_i = \eta(X_i)$ and let $b \in F^n$ be the element (b_1, \dots, b_n) . Then $f(b) = 0$ for all $f \in P$ and $g(b) \neq 0$. We would like to find such an element in K . Since ideals in polynomial rings are finitely generated, we can find polynomials f_1, \dots, f_m that generate P . The sentence

$$\exists v_1 \cdots \exists v_n (f_1(\bar{v}) = \cdots = f_m(\bar{v}) = 0 \wedge g(\bar{v}) \neq 0)$$

is true in F . Thus it is also true in K and we can find $a \in K^n$ such that each $f \in P$ vanishes at a but $g(a) \neq 0$. \square

Notice that the above proof has the same basic structure as the result about polynomial maps on \mathbb{C}^n . The idea was to come up with a different field, in this case F , where the result was easy to prove, and use logical ideas to deduce the result for the field we were originally interested in, in this case K .

5 The Reals

Quantifier elimination in the language of rings does not work in the field of real numbers. For instance, the formula

$$\exists y \, x = y \cdot y,$$

which asserts “ x is a square,” is not equivalent to a quantifier-free formula in the language of rings. Of course, x is a square if and only if $x \geq 0$. So we *could* eliminate this quantifier if we were prepared to add a symbol for the ordering to our language. An amazing result of Tarski shows that this is the only obstruction to quantifier elimination.

Let \mathcal{L}_{or} be the language of ordered rings, which is the language of rings with the addition of the symbol “ $<$ ” for an ordering. Which \mathcal{L}_{or} -sentences are true in the real field? Some of the properties of \mathbb{R} that we can formalize in \mathcal{L}_{or} include:

- (i) the axioms for ordered fields, such as the sentence

$$\forall x \forall y \, (x > 0 \wedge y > 0) \rightarrow x \cdot y > 0;$$

- (ii) the intermediate-value property for polynomials, which states that if $p(x)$ is a polynomial and there exist a and b such that $a < b$ and $p(a) < 0 < p(b)$, then there exists c such that $a < c < b$ and $p(c) = 0$.

$p(b)$, then there exists a real number c such that $a < c < b$ and $p(c) = 0$.

The intermediate-value property is expressed not by just one sentence, but by the infinite sequence of sentences

$$\forall d_0 \cdots \forall d_n \forall a \forall b \left(\sum d_i a^i < 0 < \sum d_i b^i \rightarrow \exists c \sum d_i c^i = 0 \right),$$

one for each positive integer n .

An ordered field that satisfies the intermediate-value property is called a *real closed* field. It turns out that an equivalent way of axiomatizing real closed fields is as ordered fields for which every positive element is a square and every polynomial of odd degree has a zero. Tarski’s theorem is the following statement.

Theorem. For any \mathcal{L}_{or} -formula ϕ there is a quantifier-free \mathcal{L}_{or} -formula ψ such that ϕ and ψ are equivalent in every real closed field.

What are the quantifier-free formulas of \mathcal{L}_{or} ? It turns out (and is not hard to show) that they are finite Boolean combinations of formulas of the form $p(v_1, \dots, v_n) = q(v_1, \dots, v_n)$ and formulas of the form $p(v_1, \dots, v_n) < q(v_1, \dots, v_n)$, where, as in the case of \mathcal{L}_{rng} , p and q are polynomials in n and m variables, respectively, with integer coefficients. As for quantifier-free sentences, they are Boolean combinations of sentences of the form $k = l$ and sentences of the form $k < l$.

One consequence of quantifier elimination is the following result, which tells us that every \mathcal{L}_{or} statement that is true in \mathbb{R} can be proved from the real-closed-field axioms. One says that these axioms *completely axiomatize* the theory of the real field.

Corollary. Let K be a real closed field and let ϕ be an \mathcal{L}_{or} -sentence. Then $K \models \phi$ if and only if $\mathbb{R} \models \phi$.

To prove this, first use Tarski’s theorem to find a quantifier-free sentence ψ such that ϕ and ψ are equivalent in any real closed field. Every ordered field has characteristic zero and contains the rational numbers as an ordered subfield. Therefore \mathbb{Q} is a subfield of both K and \mathbb{R} . But the very simple nature of quantifier-free sentences in \mathcal{L}_{or} means that

$$K \models \psi \Leftrightarrow \mathbb{Q} \models \psi \Leftrightarrow \mathbb{R} \models \psi.$$

Since ϕ and ψ are equivalent in all real closed fields, it follows that $K \models \phi$ if and only if $\mathbb{R} \models \phi$.

PUP: square at end of line here does indeed signify end of proof.

By the completeness theorem, ϕ is true in every real closed field if and only if we can prove ϕ from the axioms for real closed fields, and ϕ is false in every real closed field if and only if we can prove $\neg\phi$ from the axioms for real closed fields. It follows that the \mathcal{L}_{or} -theory of the real field is decidable. Indeed, if ϕ is true in \mathbb{R} , then by the corollary above, it is true in every real closed field, so it has a proof. If ϕ is false in \mathbb{R} , then $\neg\phi$ is true in \mathbb{R} , so for the same reason $\neg\phi$ has a proof. Therefore, to decide whether ϕ is true, one can search through all possible proofs from the axioms of real closed fields until one proves either ϕ or $\neg\phi$.

Let \mathcal{M} be a mathematical structure consisting of a set M and various other parts such as functions and binary operations. A subset X of M is called *definable*, with respect to some language \mathcal{L} that describes \mathcal{M} , if there is an \mathcal{L} -formula ϕ with a free variable x such that $X = \{x \in M : \phi(x)\}$. Quantifier elimination gives us a good geometric understanding of the definable sets. If K is an ordered field, we say that $X \subseteq K^n$ is *semialgebraic* if it is a finite Boolean combination of sets of the form

$$\{x \in K^n : p(x) = 0\} \quad \text{and} \quad \{x \in K^n : q(x) > 0\},$$

where $p, q \in K[X_1, \dots, X_n]$. By quantifier elimination, the definable sets in a real closed field are easily shown to be exactly the semialgebraic sets.

A simple application of this fact is that if A is a semialgebraic subset of \mathbb{R}^n , then the closure of A is also semialgebraic. Indeed, the closure of A is, by definition, the set

$$\left\{x \in \mathbb{R}^n : \forall \epsilon > 0 \exists y \in A \sum_{i=1}^n (x_i - y_i)^2 < \epsilon\right\}.$$

This is a definable set, and hence a semialgebraic set.

Semialgebraic subsets of the real line are particularly simple. For any real polynomial f in one variable, the set $\{x \in \mathbb{R} : f(x) > 0\}$ is a finite union of open intervals. Therefore, any semialgebraic subset of \mathbb{R} is a finite union of points and intervals. This simple fact is the starting point of the modern model-theoretic approach to \mathbb{R} . Let \mathcal{L}^* be a language extending \mathcal{L}_{or} and let \mathbb{R}^* denote the reals considered as an \mathcal{L}^* -structure. For example, below we will be interested in the case where $\mathcal{L}^* = \mathcal{L}_{\text{exp}}$ and $\mathbb{R}^* = \mathbb{R}_{\text{exp}}$. We say that \mathbb{R}^* is *o-minimal* if every subset of \mathbb{R} definable using \mathcal{L}^* -formulas is a finite union of points and intervals. The “o” in “o-minimal” stands for “ordered.” \mathbb{R}^* is o-minimal if every definable subset of \mathbb{R} can be defined using only the ordering.

Pillay and Steinhorn introduced o-minimality, generalizing an earlier idea of van den Dries. It turned out

to be a key definition, because although o-minimality is defined in terms of the one-dimensional set \mathbb{R} , it has remarkably strong consequences for definable subsets of \mathbb{R}^n when $n > 1$.

To explain this, we inductively define a collection of basic sets called *cells* as follows.

- A subset X of \mathbb{R} is a cell if and only if it is either a point or an interval.
- If X is a cell in \mathbb{R}^n and f is a continuous definable function from X to \mathbb{R} , then the graph of f (which is a subset of \mathbb{R}^{n+1}) is a cell.
- If X is a cell in \mathbb{R}^n and f and g are continuous definable functions from X to \mathbb{R} such that $f(x) > g(x)$ for every $x \in X$, then $\{(x, y) : x \in X \text{ and } f(x) > y > g(x)\}$ is a cell, as are $\{(x, y) : x \in X \text{ and } f(x) > y\}$ and $\{(x, y) : x \in X \text{ and } y > f(x)\}$.

Cells are topologically simple definable sets that play the role of open intervals in \mathbb{R} . It is not hard to see that any cell is homeomorphic to $(0, 1)^n$ for some n . Remarkably, all definable sets can be decomposed into cells. The following theorem is a precise version of this statement.

Theorem.

- If \mathbb{R}^* is an o-minimal structure, then every definable set X can be partitioned into finitely many disjoint cells.
- If $f : X \rightarrow \mathbb{R}$ is a definable function, then there is a partition of X into finitely many cells such that f is continuous on each cell.

This is just the beginning. In any o-minimal structure, definable sets have many of the good topological and geometric properties of the semialgebraic sets. For example:

- Any definable set has finitely many connected components.
- Definable bounded sets can be definably triangulated.
- Suppose that X is a definable subset of \mathbb{R}^{n+m} . For each $a \in \mathbb{R}^m$, let X_a be the “cross-section” $\{x \in \mathbb{R}^n : (x, a) \in X\}$. Then there are only finitely many different homeomorphism types for the sets X_a .

As these results were known for semialgebraic sets, the real interest is in finding new o-minimal structures. The most interesting example is \mathbb{R}_{exp} . It is known that

\mathbb{R}_{exp} does not have quantifier elimination in the language \mathcal{L}_{exp} . Wilkie showed that the next best thing is true. We say that \mathbb{R}^n is an *exponential variety* if it is the zero set of a finite system of exponential terms. For example, the set $\{(x, y, z) : x = \exp(y)^2 - z^3 \wedge \exp(\exp(z)) = y - x\}$ is an exponential variety.

Theorem. *Every \mathcal{L}_{exp} -definable subset of \mathbb{R}^n is of the form*

$$\{x \in \mathbb{R}^n : \exists y \in \mathbb{R}^m (x, y) \in V\}$$

for some exponential variety $V \subseteq \mathbb{R}^{n+m}$.

In other words, the definable sets, though not exponential varieties themselves, are projections of exponential varieties, which makes them tractable. Indeed, a theorem from real analytic geometry, due to Khovanskii, states that every exponential variety has a finite number of connected components. Since this property is preserved by projections, it follows that every definable set has a finite number of connected components, and also that every definable subset of the real line is a finite union of points and intervals. Thus \mathbb{R}_{exp} is o-minimal and all of the results above about definable sets in o-minimal structures apply.

Tarski asked if the theory of \mathbb{R}_{exp} is decidable. This question remains open, but the answer is known to follow from the following conjecture of Schanuel in transcendental number theory.

Conjecture. *Suppose that $\lambda_1, \dots, \lambda_n$ are complex numbers that are linearly independent over \mathbb{Q} . Then the field $\mathbb{Q}(\lambda_1, \dots, \lambda_n, e^{\lambda_1}, \dots, e^{\lambda_n})$ has transcendence degree at least n .*

Macintyre and Wilkie have shown that if Schanuel's conjecture is true, then the theory of \mathbb{R}_{exp} is decidable.

6 The Random Graph

Model-theoretic methods give interesting information about random GRAPHS [III.34]. Suppose we construct a graph as follows. The vertex set is the set \mathbb{N} of all natural numbers \mathbb{N} . To decide whether we will have an edge between x and y (with $x \neq y$) we flip a coin, putting an edge there if and only if we get heads. Although these constructions are random, we will show below that, with probability 1, any two such graphs are isomorphic.

The proof depends on the following extension property. Let A and B be disjoint finite subsets of \mathbb{N} , and suppose that they have sizes n and m , respectively. We would like to find a vertex $x \in \mathbb{N}$ that is joined to every element of A and to no element of B . Now for any

particular x , the probability that it does *not* have the desired property is $p = 1 - 2^{-(n+m)}$. Therefore, if we look at N different vertices, the probability that none of them has the desired property is p^N . Since this converges to zero with N , the probability that at least one $x \in \mathbb{N}$ has the property is 1. Moreover, since there are only countably many disjoint pairs (A, B) of finite sets, with probability 1 it is the case that for *every* such pair (A, B) one can find a vertex x that is joined to every vertex in A and to no vertex in B .

We can formalize this observation in a model-theoretic way. Let $\mathcal{L}_g = \mathcal{L}(\sim)$, where “ \sim ” is a binary relation symbol (which we read as “is joined to”). We let T be the \mathcal{L}_g -theory:

- (i) $\forall x \forall y (x \sim y \rightarrow y \sim x)$;
- (ii) $\forall x \neg(x \sim x)$;
- (iii) $\Phi_{n,m}$ for $n, m \geq 0$.

Here $\Phi_{n,m}$ is the sentence

$$\forall x_1 \cdots \forall x_n \forall y_1 \cdots \forall y_m \left(\bigwedge_{i=1}^n \bigwedge_{j=1}^m x_i \neq y_j \rightarrow \exists z \bigwedge_{i=1}^n (R(x_i, z) \wedge \neg R(y_i, z)) \right).$$

The first two sentences tell us that the relation “ \sim ” defines a graph, and for each pair (n, m) the sentence $\Phi_{n,m}$ tells us that the extension property holds for all pairs of disjoint sets A and B with A of size n and B of size m . Thus, a model of T is a graph for which the extension property holds for any pair of disjoint finite sets of vertices.

The argument above shows that with probability 1 the random graphs we constructed are models of T . Now let us see why they are isomorphic (again with probability 1). This will be an immediate consequence of the following theorem.

Theorem. *If G_1 and G_2 are any two countable models of T , then G_1 is isomorphic to G_2 .*

Recall that an *isomorphism* between G_1 and G_2 means a bijection f from the vertex set of G_1 to the vertex set of G_2 such that x is joined to y in G_1 if and only if $f(x)$ is joined to $f(y)$ in G_2 . The proof, which we shall now sketch, is a “back-and-forth” argument that gradually builds up an isomorphism between G_1 and G_2 . First, let a_0, a_1, \dots be an enumeration of the vertices of G_1 and let b_0, b_1, \dots be an enumeration of the vertices of G_2 . Let us set $f(a_0)$ to be b_0 . Next, we choose an image for a_1 : if a_1 is joined to a_0 then we need to find some vertex that is joined to b_0 and if a_1 is not joined to a_0

then we need to find a vertex that is not joined to b_0 . Either way, we can do it because G is a model of T , so it satisfies the extension property. (The particular cases we use here are $\Phi_{1,0}$ and $\Phi_{0,1}$.)

It is tempting to continue finding images for a_2, a_3 , and so on, in each case using the extension property to make sure that the images are joined to each other if and only if the original vertices are. The trouble with this is that we may not end up with a bijection, since for any particular b_j there is no guarantee that we will ever choose it as the image of some a_j . However, we can remedy this by alternately choosing an image for the first a_i that does not yet have an image, and a preimage for the first b_j that does not yet have a preimage. In this way we build the desired isomorphism.

It was not essential to use model theory to prove the above result. However, it has the following very nice model-theoretic consequence.

Corollary. *For any \mathcal{L}_g -sentence ϕ either ϕ is true in every model of T or $\neg\phi$ is true in every model of T . Moreover, there is an algorithm that will tell us which of ϕ or $\neg\phi$ is true in every model of T .*

To prove this, one first applies a slight strengthening of the compactness theorem, which allows one to conclude that if the result is false then there are *countable* models G_1 and G_2 of T such that ϕ is true in G_1 and $\neg\phi$ is true in G_2 . But this shows that G_1 and G_2 are not isomorphic, and therefore directly contradicts the previous theorem.

To decide which of ϕ or $\neg\phi$ is true in every model of T , one searches through all possible proofs from the sentences of T . By the completeness theorem, one or other of the statements has a proof, so we will eventually find either a proof of ϕ or a proof of $\neg\phi$. At that point we will know which of ϕ and $\neg\phi$ is true in every model of T .

The theory T also gives us information about random finite graphs. Let \mathcal{G}_N be the set of all graphs with vertices $\{1, 2, \dots, N\}$. We consider the probability measure on \mathcal{G}_N in which we make all graphs equally likely. This is the same as constructing a random graph on N vertices, where for each i and j we toss an unbiased coin in order to decide whether i is joined to j . For any \mathcal{L}_g -sentence ϕ , let us write $p_N(\phi)$ for the probability that a random graph on N vertices satisfies ϕ .

An easy variant of the argument for infinite graphs shows that for each extension axiom $\Phi_{n,m}$, the probability $p_N(\Phi_{n,m})$ tends to 1. Therefore, for any fixed M ,

if N is sufficiently large, then with very high probability a random graph on N vertices satisfies all the axioms $\Phi_{n,m}$ with $n, m \leq M$.

This observation allows us to use the theory T to get a good understanding of the asymptotic properties of random graphs. The following result is called a *zero-one law*.

Theorem. *For any \mathcal{L}_g -sentence ϕ , the probability $p_N(\phi)$ either tends to 0 or tends to 1 as $N \rightarrow \infty$. Moreover, T axiomatizes the set of statements ϕ such that the limit is 1, called the almost sure theory of graphs, which is a decidable theory.*

This follows from our previous results. We saw earlier that either ϕ is true in every model of T or $\neg\phi$ is true in every model of T . In the first case, by the completeness theorem there must be a proof of ϕ from T . Since proofs are finite, this proof can use only finitely many of the statements $\Phi_{n,m}$. Therefore, there exists some M such that if $G \models \Phi_{M,M}$, then $G \models \phi$. But if G is a random graph on N vertices, then the probability that $G \models \Phi_{M,M}$ tends to 1, and therefore the probability $p_N(\phi)$ that $G \models \phi$ tends to 1 as well. The same argument holds if $\neg\phi$ is true in every model of T and shows that $p_N(\neg\phi)$ tends to 1, which implies that $p_N(\phi)$ tends to 0.

Note the following interesting consequence of this result. It is not hard to prove that the probability that a random graph contains at least $\frac{1}{2} \binom{N}{2}$ edges converges to $\frac{1}{2}$ as N tends to infinity. Combining this simple observation with the theorem we can deduce that the property “contains at least as many edges as nonedges” cannot be expressed by a first-order formula in \mathcal{L}_g . This is a purely syntactic result, but to prove it we made essential use of model theory.

Further Reading

Shoenfield (2001) is an excellent introduction to logic including the completeness and incompleteness theorems, basic computability theory, and elementary model theory.

The examples described here give only a small part of the flavor for modern model theory. Hodges (1993), Marker (2002), and Poizat (2000) are comprehensive introductions. Marker et al. (1995) contains several introductory articles on the model theory of fields.

In addition to providing tools for analyzing definability in particular structures, a major goal in model theory is proving structure theorems for wide classes

of mathematical structures. A key feature is the development by Shelah of notions of dependence generalizing linear dependence in vector spaces and algebraic dependence in fields. Led by Hrushovski and Zilber, model theorists have studied the geometry of dependence and found that frequently it can be used to detect hidden algebraic structure.

In recent years, the abstract model theory has found interesting applications in classical mathematics. Hrushovski used these ideas to give a model-theoretic proof of the Mordell–Lang conjecture for function fields in Diophantine geometry. Bouscaren (1998) is an excellent collection of survey articles leading up to Hrushovski’s proof.

Bouscaren, E., ed. 1998. *Model Theory and Algebraic Geometry. An Introduction to E. Hrushovski’s Proof of the Geometric Mordell–Lang Conjecture*. New York: Springer.

Hodges, W. 1993. *Model Theory*. Encyclopedia of Mathematics and Its Applications, volume 42. Cambridge: Cambridge University Press.

Marker, D. 2002. *Model Theory: An Introduction*. New York: Springer.

Marker, D., M. Messmer, and A. Pillay. 1995. *Model Theory of Fields*. New York: Springer.

Poizat, B. 2000. *A Course in Model Theory. An Introduction to Contemporary Mathematical Logic*. New York: Springer.

Shoenfield, J. 2001. *Mathematical Logic*. Natick, MA: A. K. Peters.

IV.24 Stochastic Processes

Jean-François Le Gall

1 Historical Introduction

Stochastic processes are one of the major themes of modern probability theory. Roughly speaking, they are mathematical models that describe the evolution of random phenomena as time goes by. In this article, we shall introduce and illustrate the fundamental ideas of the theory of stochastic processes by concentrating on the single most important example: Brownian motion. We start with a brief historical introduction, in order to provide some motivation for the mathematical theory that follows.

In 1828, the British botanist Robert Brown observed the very irregular and wiggly motion of small particles of pollen suspended in water. Brown pointed out the unpredictable character of the motion, which appeared to obey no known physical rule. During the nineteenth century, several physicists tried to understand the origin of this “Brownian motion,” which turned out to

be present in many other physical phenomena. Several theories were proposed, some of them rather fanciful: perhaps Brownian particles were living microscopic animals, or perhaps the motion was due to electrostatic forces. By the end of the century, however, physicists had concluded that the constant changes of direction in Brownian motion could be explained by the impacts on a particle from the molecules of the surrounding medium. If the particle was sufficiently light, then these numerous collisions could have a macroscopic influence on its displacement. This explanation was also consistent with the experimental observation that the motion became faster if the temperature of the water, and thus the thermal agitation of its molecules, increased.

Albert Einstein, in one of his three famous 1905 papers, was responsible for a major step forward in the understanding of Brownian motion. He worked out that if a Brownian particle starts at the origin, then after a fixed time t its position should be randomly distributed according to the (three-dimensional) GAUSSIAN DISTRIBUTION [III.73 §5] with mean 0 and variance $\sigma^2 t$, where σ^2 is a constant, called the *diffusion constant*, that measures how quickly the distribution spreads out with time. (One can think of this loosely as the speed of the Brownian motion, but we shall see later that the word “speed” is not really appropriate.) Einstein’s method was based on considerations of statistical physics, which led him to THE HEAT EQUATION [I.3 §5.4] and then to the Gaussian density that solves this equation (see section 5.2).

A few years before Einstein, the French mathematician Louis Bachelier, in his work about the mathematical modeling of stock markets, had already noticed the Gaussian distribution of Brownian motion. However, Bachelier was dealing not with the physical phenomenon known as Brownian motion, but rather with random walks where the step size was very small. As we shall see in sections 2 and 3, the two concepts are essentially equivalent from a mathematical viewpoint. Bachelier pointed out what we call today the *Markov property* of Brownian motion: if we wish to predict the displacement after time t of a Brownian particle, then knowledge of the path followed by the particle before time t does not help us any more than just knowing the position at time t . Bachelier’s arguments were not completely satisfactory, and his ideas were not fully appreciated in his time.

How does one go about modeling a particle that moves in a random way? A first remark is that the posi-

tion of the particle at time t will be a RANDOM VARIABLE [III.73 §4] B_t . But these random variables will depend on each other: if you know where the particle is at time t , it will affect your knowledge of how likely it is to be in a certain region at some later time. These two considerations can be accommodated if we take as our basic model a set of random variables B_t , one for each non-negative real number, all defined on the same underlying probability space. This, formally speaking, is what a stochastic process is.

This may seem a rather simple definition, but in order for a stochastic process to be interesting it needs to have additional properties, and difficult mathematical questions arise as soon as one tries to obtain them. Let us write Ω for the underlying probability space. Then each of the random variables B_t is a function from Ω to \mathbb{R}^3 , and therefore we associate a point in \mathbb{R}^3 with each pair (t, ω) (where t is a positive real number and ω belongs to Ω). So far we have thought about the probability distribution of B_t , so we have been focusing on what happens when we fix t and let ω vary. However, we must also consider what happens when we look at a “single instance” of a stochastic process, by fixing ω and letting t vary. For fixed ω , the function that takes t to $B_t(\omega)$ is called a *sample path*. If we want a rigorous mathematical theory of Brownian motion, then a very important property it should satisfy is that all the sample paths are continuous: that is, for fixed ω the point $B_t(\omega)$ depends continuously on t .

Physical observations, as well as the contributions of Einstein and Bachelier described above, suggested a few other properties that Brownian motion should satisfy. It then became a substantial mathematical problem to prove that there existed a stochastic process with those properties. Wiener was the first person to establish this, which he did in 1923, and for this reason the mathematical concept of Brownian motion is sometimes called the *Wiener process*.

The most famous names of probability theory in the twentieth century, including KOLMOGOROV [VI.88], Lévy, Itô, and Doob, all made important contributions to the study of Brownian motion. Detailed properties of the sample paths have received particular attention, ever since the physicist Jean Perrin observed that these functions are nowhere differentiable (despite Wiener’s later result that they were continuous). The nondifferentiability of Brownian trajectories led Itô to introduce a differential calculus for functions of Brownian motion and more general stochastic processes. This Itô stochastic calculus, which will be briefly presented

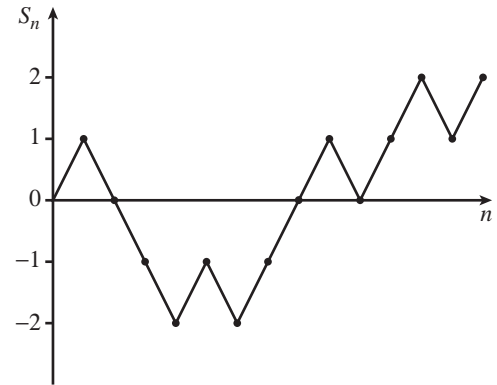


Figure 1 The accumulated gain in coin tossing.

in section 4, has found many applications in many different areas of modern probability theory.

2 Coin Tossing and Random Walks

One of the easiest ways to understand Brownian motion is via another important concept of probability: that of *random walks*. Suppose you were to play a game where you repeatedly tossed a coin, winning €1 if it came up heads, and losing €1 if it came up tails. One could then define a sequence of random variables S_0, S_1, S_2, \dots , where S_n represented your total gain (which could well be negative) after n tosses of the coin. Two simple properties of this sequence are that S_0 must be 0 and that S_n and S_{n-1} always differ by 1. One can see this in figure 1, which plots a graph of the sequence in the case where the coin tosses are HTTHTHTHHHTHTH....

A third property becomes clear if one defines another sequence of random variables $\varepsilon_1, \varepsilon_2, \dots$, representing the outcome of each toss of the coin. These are independent, and each ε_n takes the value 1 with probability $\frac{1}{2}$ and -1 with probability $\frac{1}{2}$. Moreover, for each n we can write $S_n = \varepsilon_1 + \dots + \varepsilon_n$. The distribution of sums of this kind depends in a very simple way on the well-known BINOMIAL DISTRIBUTION [III.73 §1]. (To be precise, the binomial distribution tells you that the probability that the number of heads after n tosses is k is $2^{-n} \binom{n}{k}$. If it is k , then $S_n = k - (n - k) = 2k - n$.) What is more, if $m > 0$ then $S_{m+n} - S_m = \varepsilon_{m+1} + \dots + \varepsilon_{m+n}$, which is also a sum of n of the ε_i , so the distribution of $S_{m+n} - S_m$ is the same as that of S_n . Note too that it is independent of the values of S_0, S_1, \dots, S_m .

The name “random walk” comes from the fact that we can think of the sequence S_0, S_1, S_2, \dots as taking a succession of random steps, each of either 1 or -1 .

Brownian motion can be thought of as the limit of this process as the number of steps gets larger and larger and the sizes of the steps get correspondingly smaller.

To see what “correspondingly” means here, we appeal to the CENTRAL LIMIT THEOREM [III.73 §5], which tells us about the limiting behavior of the distribution of S_n when n gets large. Or rather, it tells us about the distribution of $(1/\sqrt{n})S_n$: the reason it is appropriate to divide by \sqrt{n} is that \sqrt{n} is the STANDARD DEVIATION [III.73 §4] of S_n . This one can think of as its “typical size”: thus, when we divide by it, the “renormalized” distribution will have “typical size” 1 (and therefore we will get the same typical size for each n).

The precise information that the central limit theorem gives us is that for any real numbers a and b with $a < b$, the probability that $a < (1/\sqrt{n})S_n < b$ tends to

$$\frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx$$

as n tends to ∞ . That is, the limiting behavior of the distribution of $(1/\sqrt{n})S_n$ is Gaussian with mean 0 and standard deviation 1. Since the distribution of $S_{m+n} - S_m$ is the same as that of S_n (as we saw earlier), this also tells us the limiting behavior of the distribution of $(1/\sqrt{n})(S_{m+n} - S_m)$ for any m .

3 From Random Walks to Brownian Motion

In the previous section, we looked at a sequence of random variables S_0, S_1, S_2, \dots . This is another stochastic process, except that “time” is now represented by a positive integer. (One says that it is a *discrete-time* process.) Now let us try to do justice to the idea that Brownian motion is something like a random walk with infinitely many infinitesimally small steps. (We are now looking at one-dimensional Brownian motion, rather than the three-dimensional Brownian motion discussed right at the beginning of this article.)

It will be slightly simpler to think about a Brownian motion B_t that runs just for times t between 0 and 1. We hope that the distributions of B_t , and in particular of B_1 , will be Gaussian, and the results from the last section suggest that this is exactly what we should expect if they are appropriately scaled limits of the distributions of the S_n . To be precise, suppose we have a graph like that of figure 1 but with some large number of steps n . Then the x -axis will go from 1 to n and the standard deviation of the height of the end of the graph will be \sqrt{n} . Therefore, if we shrink the graph horizontally by a factor of n and vertically by a factor of \sqrt{n} we will obtain the graph of a random function $S^{(n)}$ from $[0, 1]$

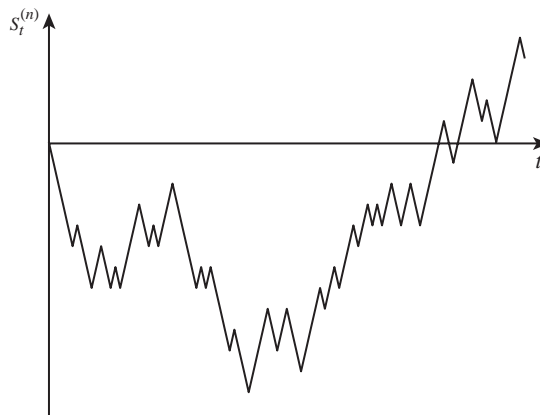


Figure 2 The rescaled random walk $S^{(n)}$ for $n = 100$.

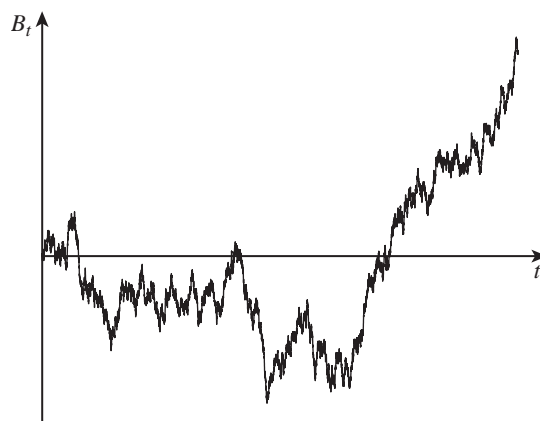


Figure 3 Simulation of linear Brownian motion.

to \mathbb{R} , and the standard deviation of $S^{(n)}(1)$ will be 1. Effectively, we are shrinking the time between the steps of the random walk from 1 to $1/n$ and shrinking the step size from 1 to $1/\sqrt{n}$. Also, so that the functions $S^{(n)}$ are defined everywhere, we “join the dots” of the graph with straight lines, just as we did in figure 1. A rescaled random walk of this kind is shown in figure 2.

At this point, we shall simply assume that the distributions of these rescaled random walks converge, in an appropriate sense, to a stochastic process with continuous sample paths. This stochastic process is the Brownian motion B_t . The graph of a typical sample path is illustrated in figure 3. Notice how similar its general behavior is to that of the graph in figure 2.

If we want to approximate a Brownian motion that goes on forever rather than stopping at 1, all we have

to do is let the rescaled random walk go on forever, rather than stopping after n steps.

Now let us give a more precise definition. A *linear Brownian motion starting at x* is a collection $(B_t)_{t \geq 0}$ of real-valued random variables with the following properties.

- $B_0 = x$. (In other words, $B_0(\omega) = x$ for every ω in the underlying probability space.)
- The sample paths are continuous.
- Given any $s < t$ the distribution of $B_t - B_s$ is Gaussian with mean 0 and variance $t - s$.
- Moreover, $B_t - B_s$ is independent of the process up to time s . (This implies the Markov property mentioned in section 1.)

Each of these properties has its counterpart for random walks, as we saw in the previous section. Therefore, even though it is not easy to prove that Brownian motion exists, the result is nevertheless highly plausible. (It turns out to be easy to construct a stochastic process that satisfies all the properties above apart from the second; the difficulty is in obtaining the continuity of the sample paths.) Another important remark is that the above properties characterize Brownian motion: any two stochastic processes with those properties are essentially the same.

We have not yet said what it means for the rescaled random walks $S^{(n)}$ to “converge” to Brownian motion. Rather than defining this notion precisely, we shall merely remark that any “reasonable” function that we can define on the processes $S^{(n)}$ will converge to the “corresponding” function of the limiting Brownian motion B_t . For example, as we have already seen, the probability that $S^{(n)}(1)$ lies between a and b converges to

$$\frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx.$$

But B_1 is governed by the Gaussian distribution, so this is also the probability that B_1 lies between a and b .

A more interesting example is the proportion X_n of times t between 0 and 1 for which $S^{(n)}(t)$ is positive, or rather the way that this proportion (which is a random variable that depends on the walk $S^{(n)}$) is distributed. This “converges in distribution” to the distribution of the corresponding proportion X for Brownian motion. That is, for any $a < b$, the probability that the proportion X_n lies between a and b converges to the probability that the proportion X lies between a and b . The probability distribution for X is known explicitly, and

is called *Paul Lévy’s arcsine law*:

$$P[a \leq X \leq b] = \int_a^b \frac{dx}{\pi \sqrt{x(1-x)}}.$$

Perhaps surprisingly, X is more likely to be close to 0 or 1 than to $\frac{1}{2}$. The basic reason for this is that if s and t are two different times, then the events $B_s > 0$ and $B_t > 0$ are positively correlated.

The convergence of random walks to Brownian motion is just one special case of a much more general phenomenon (see, for example, Billingsley 1968). For instance, we can allow other probability distributions for the individual steps of the random walk. A typical result is that if each individual step has mean 0 (as is the case when we have $+1$ or -1 with probability $\frac{1}{2}$) and finite variance, then the limiting process will always be a simple rescaling of Brownian motion. In this sense Brownian motion appears as a universal object: it is the continuous limit of a wide range of discrete models. (See the introduction to PROBABILISTIC MODELS OF CRITICAL PHENOMENA [IV.25] for a discussion of universality.)

Now that we have discussed one-dimensional Brownian motion, let us think about how to model random continuous paths in three dimensions. An obvious way of doing it would be to take three independent Brownian motions, B_t^1 , B_t^2 , and B_t^3 , and let these be the three coordinates of a point in a random path in \mathbb{R}^3 . And indeed, this is how three-dimensional Brownian motion is defined. However, it is not quite so obvious that this is a good definition. In particular, it seems to depend on our choice of coordinate system, which is worrying if we want a good model for physical Brownian motion.

However, a key property of higher-dimensional Brownian motion (the definition just given clearly generalizes to any dimension d) is *rotational invariance*. That is, if we choose a different ORTHONORMAL BASIS [III.37] as our coordinate system, then we obtain the same stochastic process. The proof of this is a simple deduction from the basic fact that the DENSITY FUNCTION [III.73 §3] of a vector made up of d independent one-dimensional Gaussian random variables is

$$\frac{1}{(2\pi)^{d/2}} e^{-(x_1^2 + \dots + x_d^2)/2}.$$

Since the quantity $x_1^2 + \dots + x_d^2$ is just the square of the distance from 0 to (x_1, \dots, x_d) , the density does not change when you rotate.

In the planar case $d = 2$, there is a much deeper invariance property, which we shall explain in section 5.3.

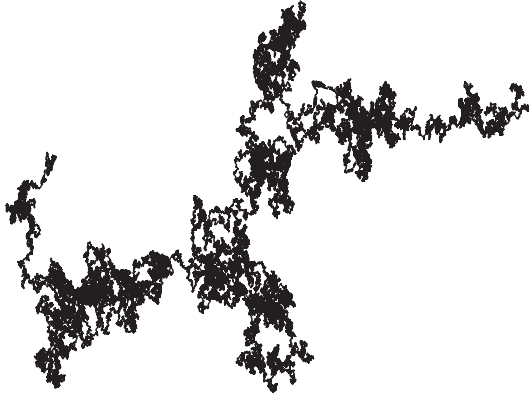


Figure 4 Simulation of planar Brownian motion.

It is not hard to incorporate the notion of a diffusion constant into our model. (This is the constant σ^2 mentioned in section 1 that measures how quickly the Brownian motion tends to spread out.) All one has to do is rescale from B_t to $B_{\sigma^2 t}$.

As one might expect, higher-dimensional Brownian motions are limits of higher-dimensional random walks. This helps to explain why mathematical Brownian motion is a good model for the physical phenomenon observed by Brown: the erratic displacements caused by collisions with molecules resemble the steps of a random walk with very small step size. See figure 4 for a simulation of the curve of a planar Brownian motion over the time interval $[0, 1]$.

4 Itô's Formula and Martingales

Let f be a real-valued differentiable function. Suppose that we are told the values of $f'(x)$ at $0, 1/n, 2/n, \dots, (n-1)/n$ for some large positive integer n and are asked to estimate $f(1) - f(0)$. If the derivative f' did not vary too rapidly, then we would expect the difference $f((j+1)/n) - f(j/n)$ to be approximately $(1/n)f'(j/n)$, so a good approximation ought to be

$$\frac{1}{n} \left(f'(0) + f'\left(\frac{1}{n}\right) + f'\left(\frac{2}{n}\right) + \dots + f'\left(\frac{n-1}{n}\right) \right).$$

THE FUNDAMENTAL THEOREM OF CALCULUS [I.3 §5.5] implies that this argument is indeed correct if the derivative f' is continuous.

Now let us look at a setup that is superficially similar. This time, let us suppose that the numbers $x_0, x_1, x_2, \dots, x_n$ are the positions of a random walk with step size $1/\sqrt{n}$. Suppose that f is a function with a well-behaved derivative, and that we know the val-

ues of $f'(x)$ at x_0, x_1, \dots, x_{n-1} . This time, let us think about estimating $f(x_n) - f(x_0)$.

If we follow the lines of our previous argument, then we will comment that $f(x_{j+1}) - f(x_j)$ is approximately $(x_{j+1} - x_j)f'(x_j)$, which would lead to an estimate of

$$(x_1 - x_0)f'(x_0) + (x_2 - x_1)f'(x_1) + \dots + (x_n - x_{n-1})f'(x_{n-1}).$$

Now it is not obvious that this will still be a good estimate. The reason is that, typically, the random walk will have gone backwards and forwards, covering the same ground several times before reaching its eventual destination x_n , and this gives the errors in the approximations a chance to accumulate. To see that this is a serious problem, consider the very well-behaved function $f(x) = x^2$ and let $x_0 = 0$. In this case,

$$f(x_{j+1}) - f(x_j) = x_{j+1}^2 - x_j^2$$

and a simple calculation shows that this is equal to

$$(x_{j+1} - x_j)2x_j + (x_{j+1} - x_j)^2.$$

The first term here equals $(x_{j+1} - x_j)f'(x_j)$ and is therefore the approximation that we are considering, so the error we have to worry about is $(x_{j+1} - x_j)^2$, which is the square of the step size of the random walk. In other words, it is $1/n$. But there are n steps to the walk, so the total error (all of which is positive) is 1. Since the order of magnitude of x_n , and hence x_n^2 , is typically about 1, this is a significant fraction of $f(x_n) - f(x_0)$, and therefore our estimate is not a good one.

Remarkably, this turns out to be the “only” problem that can occur, and we can get around it rather easily. All we have to do is use one more term in the Taylor expansion. That is, we use the slightly more refined approximation

$$f(x_{j+1}) - f(x_j) = (x_{j+1} - x_j)f'(x_j) + \frac{1}{2}(x_{j+1} - x_j)^2 f''(x_j).$$

(Of course, now we are assuming that the *second* derivative f'' exists and is continuous.) Notice that in the example $f(x) = x^2$ just considered, $f''(x) = 2$ for every x , and so if we add up all the above approximations we get exactly the right answer. In general, as this observation would suggest, one can show that $f(x_n) - f(x_0)$ is well-approximated by

$$\sum_{j=0}^{n-1} (x_{j+1} - x_j)f'(x_j) + \frac{1}{2} \sum_{j=0}^{n-1} (x_{j+1} - x_j)^2 f''(x_j).$$

Now let us think about what happens to these two sums if we allow our random walks to converge to a Brownian motion B_t . A relatively straightforward argument, based on the fact that $(x_{j+1} - x_j)^2$ is just the reciprocal of the number of steps, shows that the limiting distribution of the second sum exists and is given by the integral $\frac{1}{2} \int_0^t f''(B_s) ds$. This suggests that the first sum should also converge to a limit, which indeed it does: the limit is called the *stochastic integral* and is written $\int_0^t f'(B_s) dB_s$. More precisely, one ends up with the formula

$$f(B_t) = f(B_0) + \int_0^t f'(B_s) dB_s + \frac{1}{2} \int_0^t f''(B_s) ds, \quad (1)$$

which is known as *Itô's formula*. Note the similarity to the fundamental theorem of calculus. The main difference is the extra term involving the second derivative, the so-called *Itô term*.

Why, one might wonder, is this interesting? If we wish to estimate the difference between two values of a function by integrating its derivative, why not choose a smooth path rather than a very wiggly one? The point, however, is that we are not interested in just one path. For any fixed sample path, the two sides of the above formula are just numbers, but if we think of B_t as a random variable, then they too become random variables. And since both sides are defined for all $t \geq 0$, they are actually stochastic processes. So what we are discussing is a way of integrating one stochastic process to produce another.

The reason Itô's formula is so useful is that stochastic integrals have properties that allow one to prove many facts about them. In particular, if we view the stochastic integral $\int_0^t f'(B_s) dB_s$ as a collection of random variables indexed by the parameter t , then we have a stochastic process of an especially nice sort called a *martingale*. A martingale is a stochastic process $(M_t)_{t \geq 0}$ with the property that, whenever $s \leq t$, the expected value of M_t , conditional on the values of M_r for all $r \leq s$, is just M_s .

Brownian motion is a particularly simple kind of martingale, but martingales are much more general because $M_t - M_s$ is not *independent* of the values of M_r with $r \leq s$: all one knows is that the expectation of $M_t - M_s$, given those values, is zero. Here is an example that illustrates the difference: start running Brownian motion at 0; when it first reaches 1 (if it ever does), continue with Brownian motion but at double the speed (or rather, double the diffusion constant). In this case, the behavior of $M_t - M_s$ certainly depends on what has happened up to s , but its expectation is nevertheless zero.

In a certain sense, the stochastic integral term in Itô's formula behaves like a Brownian motion "run at a varying speed," rather like the example just given. The precise result is that there exists another Brownian motion $\beta = (\beta_t)_{t \geq 0}$ such that, for every $t \geq 0$,

$$\int_0^t f'(B_s) dB_s = \beta_{\int_0^t f'(B_s)^2 ds}.$$

This is in fact true for any continuous martingale—not just one given by a stochastic integral—and the relevant time change is a quantity called the *quadratic variation* of the martingale. Therefore, the graph of a continuous martingale is obtained from that of a Brownian motion by a time-change operation. This is why Brownian motion is such a central example, and why it is important to understand its behavior before going on to deal with more general stochastic processes.

It is straightforward to generalize the previous derivation of Itô's formula to multidimensional Brownian motion. If $x = (x_1, \dots, x_d)$ and $y = (y_1, \dots, y_d)$ belong to \mathbb{R}^d and are close together, then the first approximation to $f(x) - f(y)$ is now

$$\sum_{i=1}^d (x_i - y_i) \partial_i f(y),$$

where $\partial_i f(y)$ denotes the i th partial derivative of f , evaluated at y . The vector of partial derivatives at y is usually denoted $\nabla f(y)$. It is called the *gradient* of f at y (or "grad f " for short). As for the second derivative of f , it naturally generalizes to the Laplacian Δf (for reasons that are explained in SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3 §5.4]), and we therefore arrive at the formula

$$f(B_t) = f(B_0) + \int_0^t \nabla f(B_s) \cdot dB_s + \frac{1}{2} \int_0^t \Delta f(B_s) ds.$$

The stochastic integral term is defined formally in terms of one-dimensional stochastic integrals in the obvious way:

$$\int_0^t \nabla f(B_s) \cdot dB_s = \sum_{j=1}^d \int_0^t \frac{\partial f}{\partial x_j}(B_s) dB_s^j.$$

Since stochastic integrals are martingales, the stochastic process

$$M_t^f = f(B_t) - \frac{1}{2} \int_0^t \Delta f(B_s) ds$$

is (under appropriate conditions on f) a martingale. This observation leads to the *martingale problem* for Brownian motion. To state a martingale problem for a stochastic process $(X_t)_{t \geq 0}$ is to give a collection of martingales defined as functionals of this stochastic

process—just as M^f above is defined as a certain function of $(B_s)_{s \geq 0}$. The martingale problem is said to be *well-posed* if it characterizes the distribution of the given stochastic process. In the preceding example, the martingale problem is well-posed: if we know nothing about the distribution of the process $(B_t)_{t \geq 0}$ apart from the fact that M_t^f is a martingale for every (twice continuously differentiable) function f , we can infer that B must be a Brownian motion.

Martingale problems play a fundamental role in modern probability theory (see in particular Stroock and Varadhan (1979), and also THE MATHEMATICS OF MONEY [VII.9 §2.3]). The introduction of a suitable martingale problem is often the most convenient way to specify a stochastic process, or more precisely to characterize its probability distribution.

5 Brownian Motion and Analysis

5.1 Harmonic Functions

A continuous function h defined on an open subset U of \mathbb{R}^d is called *harmonic* if the average value of h over any closed ball contained in U , or equivalently the average value over the boundary of any such ball, is equal to its value at the center of the ball. A basic result of analysis is that h is harmonic if and only if it is twice continuously differentiable and $\Delta h = 0$. Harmonic functions play an important role in several areas of mathematics as well as in physics. For instance, the electrical potential of a conductor in equilibrium is a harmonic function outside the conductor. And if the temperature of the boundary of a body is kept fixed (that is, although different parts of the boundary may have different temperatures, these temperatures do not change over time), then the equilibrium temperature inside the body is also a harmonic function. (See the discussion of the heat equation in the next section.)

Harmonic functions have a very close relationship with Brownian motion, which leads to one of the most important connections between probability and analysis. This connection is already apparent from the fact that M_t^f , defined in the previous section, is a martingale. It follows from this that $h(B_t)$ is a martingale if (and in fact only if) h is harmonic, since then the second term vanishes. However, we will explain the link between Brownian motion and harmonic functions in a more elementary way, from the classical *Dirichlet problem*. Let U be a bounded open set, and let g be a continuous real-valued function defined on the boundary ∂U

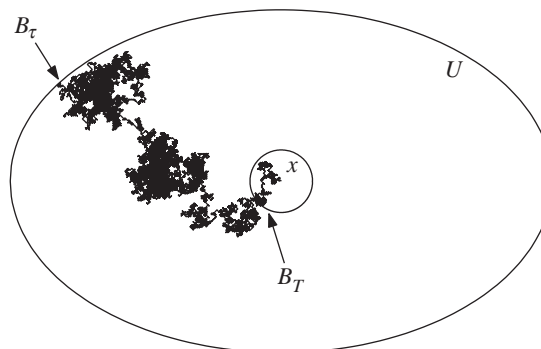


Figure 5 The probabilistic solution of the Dirichlet problem.

of U . The classical Dirichlet problem is to find a function h that is harmonic on U and is equal to g on the boundary.

The Dirichlet problem has a remarkably simple solution in terms of Brownian motion: take $x \in U$, start a Brownian motion from x , and evaluate g at the point B_T where this Brownian motion leaves U (see figure 5); then define $h(x)$ to be the average value you get. Why does this work? That is, why is the function h , defined in this way, harmonic, and why does it equal (or, to be more accurate, converge to) g at the boundary?

The answer to the last question is roughly that if x is very close to the boundary, then a Brownian motion started at x is very likely to leave U at a point close to x . Therefore, since g is continuous, the average value of g at the first exit point will be close to the value of g at any point near x .

To show that h is harmonic is more interesting. Let x be a point in U and suppose that the ball of radius r about x is contained in U . We would like to show that $h(x)$ equals the average value of h on the boundary of this ball. Now $h(x)$ is the average value of g at the point where a Brownian motion that starts at x leaves U . Let us work out this average by conditioning on the first point B_T where the Brownian path leaves the ball of radius r (see figure 5). By the rotational invariance of Brownian motion, this point will be evenly distributed around the boundary of this ball. If we reach the boundary at a point y , then the average value of g when the path leaves U (conditioning on this extra information) is $h(y)$, by definition. Therefore, $h(x)$ is indeed the average value of h on the boundary of the ball of radius r .

Convincing though this argument might seem, there is a subtlety concealed within it, connected with the

fact that a Brownian path will typically cross the boundary of the ball many times. Suppose we tried a similar argument, but this time we conditioned on the value at the *last* point where the path left the ball. If this point was y , we could not then say that the expected value of g where the path first reached the boundary of U was $h(y)$ because from that point onward the path would be forbidden to enter the ball again, and would therefore not be a Brownian motion.

Recall that the Markov property of a Brownian motion states that, given a fixed time T and another time t with $T < t$, the value of $B_t - B_T$ is independent of B_s for $s \leq T$. It may seem that we are applying this principle in the argument above, taking T to be the first time that the Brownian motion reaches the boundary of the ball. But if we do that, then T is not a fixed time since it depends on the Brownian motion. However, the argument can still be made to work because T is a so-called *stopping time*. Informally, this means that T does not depend on what the Brownian motion does after T . (Therefore the last time it leaves the ball of radius r is not a stopping time, because whether or not a given time is this last time depends on the subsequent behavior of the Brownian motion.) Brownian motion can be shown to have the *strong Markov property*, which is like the usual Markov property except that T is allowed to be a stopping time. Given this fact, it is not hard to show rigorously that h is harmonic.

5.2 The Heat Equation

Let f be a function on \mathbb{R}^d (which we shall assume to be continuous and bounded). If we think of f as a temperature distribution at time 0, then the HEAT EQUATION [III.36] models what happens to the temperature at subsequent times. To find a solution to this equation with initial value f means to find a continuous function $u(t, x)$, defined for every $t \geq 0$ and $x \in \mathbb{R}^d$, that solves the partial differential equation

$$\frac{\partial u}{\partial t} = \frac{1}{2} \Delta u \quad (2)$$

whenever $t > 0$, and that satisfies the condition $u(0, x) = f(x)$ for every x . (The factor $\frac{1}{2}$ in this equation is not important but it makes the probabilistic interpretation easier to express.)

The heat equation also has a simple solution in terms of Brownian motion: $u(t, x)$ is defined to be the expected value of $f(B_t)$ when B_t is a Brownian motion that starts at x . This tells us that heat propagates like a collection of infinitesimal Brownian particles.

The preceding probabilistic representation is quite easy to derive since one can write down an explicit formula for the expectation of $f(B_t)$ in terms of the Gaussian density function. Given this formula, all we have to do is differentiate it and check that the equation is satisfied. However, the connection between Brownian motion and the heat equation is much deeper, and in many other cases there is a probabilistic representation for a solution but no explicit formula. To take one example, suppose that we want to solve the heat equation in an open set U with Dirichlet boundary conditions. This means that we specify an initial value $f(x)$ for the temperature of each point $x \in U$ and stipulate that the temperature at the boundary is kept at 0. In other words, we want to find a function $u(t, x)$ such that $u(0, x) = f(x)$ for every $x \in U$, $u(t, x) = 0$ for every time $t \geq 0$ and every x in the boundary of U , and u satisfies the heat equation inside U . In this case, the solution is obtained as follows. Run a Brownian motion (B_t) starting at x . Let $g_t = f(B_t)$ if it has not left U at any time before t , and let $g_t = 0$ otherwise. Then define $u(t, x)$ to be the expected value of g_t .

Thus, in order to obtain the solution, we had to make just a small modification to the solution of the heat equation in \mathbb{R}^d . An analytic treatment of this version of the heat equation would be much more complicated.

5.3 Holomorphic Functions

Let us now concentrate on the case $d = 2$. As usual, we identify \mathbb{R}^2 with the complex plane \mathbb{C} . Let $f = f_1 + if_2$ be a HOLOMORPHIC FUNCTION [I.3 §5.6] defined on \mathbb{C} . Then the real part f_1 and the imaginary part f_2 of f are both harmonic functions, so that $f_1(B_t)$ and $f_2(B_t)$ are martingales. More precisely, Itô's formula tells us that, for $j = 1, 2$,

$$f_j(B_t) = f_j(x) + \int_0^t \frac{\partial f_j}{\partial x_1}(B_s) dB_s^1 + \int_0^t \frac{\partial f_j}{\partial x_2}(B_s) dB_s^2,$$

since the Itô term vanishes. As we saw in section 3, each of the two processes $f_j(B_t)$ can be expressed as a time change of a linear Brownian motion β^j . However, a stronger result can also be proved, namely that the time change is the same in both cases and that the Brownian motions β^1 and β^2 are independent. This makes it possible to prove a "localized" rotational invariance, which leads to the important *conformal invariance* property of Brownian motion. Roughly speaking, this states that the image of a planar Brownian motion under a conformal (that is, angle-preserving) mapping is another planar Brownian motion run at a different speed.

6 Stochastic Differential Equations

Imagine a Brownian particle in some water. If the temperature of the water rises, then we expect there to be more collisions with faster-moving molecules; this can be modeled easily by increasing the diffusion constant. But what if the temperature in the water varied from place to place? Then the particle would be more agitated in some parts of the water than in others. And if the water was moving, with different parts moving at different speeds, then one would need to superimpose on the Brownian motion a “drift” term, to take into account that on average we would expect the particle to move with the surrounding water.

Stochastic differential equations are used to model more complicated situations like this. Let us begin by considering the one-dimensional case. Let σ and b be two functions (which we shall assume to be continuous) defined on \mathbb{R} . We think of $\sigma(x)$ as telling us the rate of diffusion at x and of $b(x)$ as the drift at x . (For the sake of a picture, one could think of $\sigma(x)$ as the local temperature at x and $b(x)$ as the velocity at x of some “one-dimensional water.”) Let (B_t) be a one-dimensional Brownian motion.

The notation used for the associated stochastic differential equation is

$$dX_t = \sigma(X_t) dB_t + b(X_t) dt. \quad (3)$$

Here (X_t) is an unknown stochastic process. The idea is that, infinitesimally speaking, its behavior is like that of a Brownian motion with diffusivity $\sigma(X_t)$ (which is the diffusivity at the point that X_t has reached) superimposed onto a linear motion at speed $b(X_t)$. More precisely, a solution to the above equation is defined to be a continuous stochastic process (X_t) that satisfies, for every $t \geq 0$, the integral equation

$$X_t = X_0 + \int_0^t \sigma(X_s) dB_s + \int_0^t b(X_s) ds.$$

Notice that if $\sigma(x) = 0$ for every x , this boils down to the ordinary differential equation $x'(t) = b(x(t))$. The stochastic integral $\int_0^t \sigma(X_s) dB_s$ is defined by approximations similar to those described in section 4. (For this to work, there are certain technical conditions that the process (X_t) must satisfy.) In fact, stochastic differential equations were Itô’s original motivation for developing stochastic integrals.

Itô proved, under suitable conditions on σ and b , that for each $x \in \mathbb{R}$ the above equation has a unique solution (X_t) that starts at x . Furthermore, this solution is a Markov process in the sense that was explained above:

the way that (X_t) evolves after time T given the value of X_T is independent of what happens before T , and is distributed in the same way as a solution of the equation that starts at X_T . In fact, it is also a strong Markov process in the sense explained in section 5.

An important example can be found in the famous BLACK-SCHOLES MODEL [VII.9 §2] of mathematical finance. In this model, the price of a share solves a stochastic differential equation of the type above with $\sigma(x) = \sigma x$ and $b(x) = bx$, where σ and b are positive constants. This is motivated by the simple idea that the price fluctuations of a share should be roughly proportional to its current value. In this context, the number σ is called the *volatility* of the share.

The previous discussion generalizes fairly easily to stochastic differential equations in higher dimensions. The solution of a d -dimensional stochastic equation (which when $d = 3$ could model the water example mentioned at the beginning of this section) is once again a strong Markov process, known as a *diffusion process*. Much of what was said earlier about the relationship between Brownian motion and partial differential equations can be generalized to diffusion processes as well. Roughly speaking, with each diffusion process one can associate a differential operator L , and this operator plays the role that the Laplacian plays for Brownian motion.

7 Random Trees

Brownian motion and more general diffusion processes appear as limits of many discrete models in probability theory, combinatorics, and statistical physics. The most striking recent example of this is given by the so-called *stochastic Loewner evolution* (commonly abbreviated to SLE) processes, which are discussed in [IV.25 §5]). These are expected to describe the asymptotic behavior of a large number of two-dimensional models, and their definition involves both linear Brownian motion and the *Loewner equation* from complex analysis. Rather than trying to give a general presentation of the relationship between Brownian motion and discrete models, in this final section we shall discuss a surprising application of Brownian motion to random trees, which can be used to describe the genealogy of a population.

The basic discrete model is the following. We start with a single “ancestor,” which we label \emptyset . Then we place a probability distribution μ on the nonnegative integers, and use this to determine the number of children the ancestor has. Then each child is assumed to

turns out to be very closely related to a linear Brownian motion.

Notice that it cannot be a Brownian motion because it exhibits some behavior that is very untypical: it begins and ends at zero and remains positive for all time. However, we can use Brownian motion in a simple way to define a notion called a *Brownian excursion*, for which the sample paths have the right shape. The rough idea is to start a linear Brownian motion at zero, draw its graph, and then pick out the part of the graph between $x = x_1$ and $x = x_2$, where x_1 is the point where it last crosses the x -axis before $x = 1$ and x_2 is the point where it first crosses the x -axis after $x = 1$. The corresponding portion of the Brownian motion will start and end at zero and not cross zero in between. We then need to rescale it so that x goes from 0 to 1 instead of from x_1 to x_2 , and we also need to rescale the height appropriately, by dividing by $1/\sqrt{x_2 - x_1}$. Also, if the path is everywhere negative between x_1 and x_2 , we simply turn it upside down to make it positive.

Aldous's theorem states that the limiting distribution of the contour function C^{θ^n} (rescaled in time by the factor $1/2n$ and in space by the factor $1/\sqrt{2n}$, like the rescaling in section 3) is a Brownian excursion. The surprising fact about this result is that it does not depend on the offspring distribution μ . Since the contour function completely determines the shape of the corresponding tree, we find that the limiting shape of a large critical Galton-Watson tree does not depend on the offspring distribution. This is another example of universality.

This result and variants of it provide a lot of useful information about the asymptotic behavior of large trees. Many interesting functions of the tree can be rewritten in terms of the contour function and by Aldous's theorem they will converge to similar functions of the Brownian excursion, whose distribution can be computed explicitly with the help of stochastic calculus. To give just one example, this technique can be used to calculate the limiting distribution of the height of the tree θ^n . Let the variance of the offspring distribution be σ , and let us define the *rescaled height* of a tree to be its original height multiplied by $\sigma/2\sqrt{n}$. The probability that this is at least x turns out to converge, as n gets large, to the quantity

$$2 \sum_{k=1}^{\infty} (4x^2 k^2 - 1) \exp(-2k^2 x^2).$$

Acknowledgments. The author is indebted to Gilles Stoltz for his help with the simulations and to Gordon Slade for his remarks on the first version of this article.

PUP: Tim suggested adding a heading to the acknowledgements section. Style OK?

Further Reading

- Aldous, D. 1993. The continuum random tree. III. *Annals of Probability* 21:248–89.
 Bachelier, L. 1900. Théorie de la spéculation. *Annales Scientifiques de l'École Normale Supérieure* (3) 17:21–86.
 Billingsley, P. 1968. *Convergence of Probability Measures*. New York: John Wiley.
 Durrett, R. 1984. *Brownian Motion and Martingales in Analysis*. Belmont, CA: Wadsworth.
 Einstein, A. 1956. *Investigations on the Theory of the Brownian Movement*. New York: Dover.
 Revuz, D., and M. Yor. 1991. *Continuous Martingales and Brownian Motion*. New York: Springer.
 Stroock, D. W., and S. R. S. Varadhan. 1979. *Multidimensional Diffusion Processes*. New York: Springer.
 Wiener, N. 1923. Differential space. *Journal of Mathematical Physics Massachusetts Institute of Technology* 2:131–74.

IV.25 Probabilistic Models of Critical Phenomena

Gordon Slade

1 Critical Phenomena

1.1 Examples

A population can explode if its birth rate exceeds its death rate, but otherwise it becomes extinct. The nature of the population's evolution depends critically on which way the balance tips between adding new members and losing old ones.

A porous rock with randomly arranged microscopic pores has water spilled on top. If there are few pores, the water will not percolate through the rock, but if there are many pores, it will. Surprisingly, there is a critical degree of porosity that exactly separates these behaviors. If the rock's porosity is below the critical value, then water cannot flow completely through the rock, but if its porosity exceeds the critical value, even slightly, then water will percolate all the way through.

A block of iron placed in a magnetic field will become magnetized. If the magnetic field is extinguished, then the iron will remain magnetized if the temperature is below the Curie temperature 770 °C (1418 °F), but not if the temperature is above this critical value. It is striking that there is a specific temperature above which the magnetization of the iron does not merely remain small, but actually vanishes.

The above are three examples of *critical phenomena*. In each example, global properties of the system change abruptly as a relevant parameter (fertility, degree of porosity, or temperature) is varied through a critical value. For parameter values just below the critical value, the overall organization of the system is quite different from how it is for values just above. The sharpness of the transition is remarkable. How does it occur so suddenly?

1.2 Theory

The mathematical theory of critical phenomena is currently undergoing intense development. Intertwined with the science of *phase transitions*, it draws on ideas from probability theory and statistical physics. The theory is inherently probabilistic: each possible configuration of the system (e.g., a particular arrangement of pores in a rock, or of the magnetic states of the individual atoms in a block of iron) is assigned a probability, and the typical behavior of this ensemble of random configurations is analyzed as a function of parameters of the system (e.g., porosity or temperature).

The theory of critical phenomena is now guided to a large degree by a profound insight from physics known as *universality*, which, at present, is more of a philosophy than a mathematical theorem. The notion of universality refers to the fact that many essential features of the transition at a critical point depend on relatively few attributes of the system under consideration. In particular, simple mathematical models can capture some of the qualitative and quantitative features of critical behavior in real physical systems even if the models dramatically oversimplify the local interactions present in the real systems. This observation has helped to focus attention on particular mathematical models, among both physicists and mathematicians.

This essay discusses several models of critical phenomena that have attracted much attention from mathematicians, namely branching processes, the model of random networks known as the random graph, the percolation model, the Ising model of ferromagnetism, and the random cluster model. As well as having applications, these models are mathematically fascinating. Deep theorems have been proved, but many problems of central importance remain unsolved and tantalizing conjectures abound.

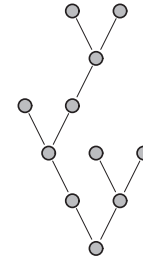


Figure 1 A possible family tree, with probability $p^{10}(1-p)^{12}$.

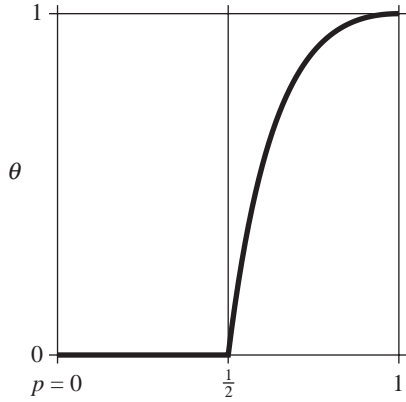
2 Branching Processes

Branching processes provide perhaps the simplest example of a phase transition. They occur naturally as a model of the random evolution of a population that changes in time as a result of births and deaths. The simplest branching process is defined as follows.

Consider an organism that lives for a unit time and that reproduces immediately before death. The organism has two potential offspring, which we can regard as the “left” offspring and the “right” offspring. At the moment of reproduction, the organism has either no offspring, a left but no right offspring, a right but no left offspring, or both a left and a right offspring. Assume that each of the potential offspring has a probability p of being born and that these two births occur independently. Here, the number p , which lies between 0 and 1, is a measure of the population’s fecundity. Suppose that we start with a single organism at time zero, and that each descendant of this organism reproduces independently in the above manner.

A possible family tree is depicted in figure 1, showing all births that occurred. In this family tree, ten offspring were produced in all, but twelve potential offspring were not born, so the probability of this particular tree occurring is $p^{10}(1-p)^{12}$.

If $p = 0$, then no offspring are born, and the family tree always consists of the original organism only. If $p = 1$, then all possible offspring are born, the family tree is the infinite binary tree, and the population always survives forever. For intermediate values of p , the population may or may not survive forever: let $\theta(p)$ denote the *survival probability*, that is, the probability that the branching process survives forever when the fecundity is set at p . How does $\theta(p)$ interpolate between the two extremes $\theta(0) = 0$ and $\theta(1) = 1$?

Figure 2 The survival probability θ versus p .

2.1 The Critical Point

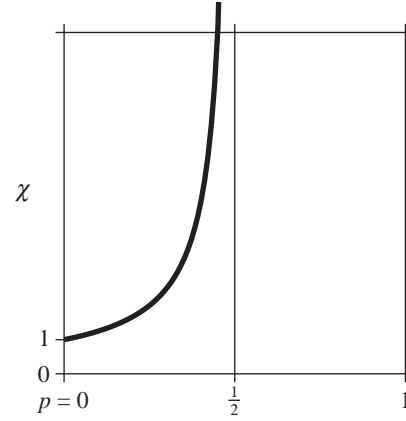
Since an organism has each of two potential offspring independently with probability p , it has, on average, $2p$ offspring. It is natural to suppose that survival for all time will not occur if $p < \frac{1}{2}$, since then each organism, on average, produces less than 1 offspring. On the other hand, if $p > \frac{1}{2}$, then, on average, organisms more than replace themselves, and it is plausible that a population explosion can lead to survival for all time.

Branching processes have a recursive nature, not present in other models, that facilitates explicit computation. Exploiting this, it is possible to show that the survival probability is given by

$$\theta(p) = \begin{cases} 0 & \text{if } p \leq \frac{1}{2}, \\ \frac{1}{p^2}(2p - 1) & \text{if } p \geq \frac{1}{2}. \end{cases}$$

The value $p = p_c = \frac{1}{2}$ is a critical value, at which the graph of $\theta(p)$ has a kink (see figure 2). The interval $p < p_c$ is referred to as *subcritical*, whereas $p > p_c$ is *supercritical*.

Rather than asking for the probability $\theta(p)$ that the initial organism has infinitely many descendants, one could ask for the probability $P_k(p)$ that the number of descendants is at least k . If there are at least $k + 1$ descendants, then there are certainly at least k , so $P_k(p)$ decreases as k increases. In the limit as k increases to infinity, $P_k(p)$ decreases to $\theta(p)$. In particular, when $p > p_c$, $P_k(p)$ approaches a positive limit as k approaches infinity, whereas $P_k(p)$ goes to zero when $p \leq p_c$. When p is strictly less than p_c , it can be shown that $P_k(p)$ goes to zero exponentially rapidly,

Figure 3 The average family size χ versus p .

but at the critical value itself we have

$$P_k(p_c) \sim \frac{2}{\sqrt{\pi k}}.$$

The symbol “ \sim ” denotes asymptotic behavior, and means that the ratio of the left- and right-hand sides in the above formula goes to 1 as k goes to infinity. In other words, $P_k(p_c)$ behaves essentially like $2/\sqrt{\pi k}$ when k is large.

There is a pronounced difference between the exponential decay of $P_k(p_c)$ for $p < p_c$ and the square-root decay at p_c . When $p = \frac{1}{4}$, family trees larger than 100 are sufficiently rare that in practical terms they do not occur: the probability is less than 10^{-14} . However, when $p = p_c$, roughly one in every ten trees will have size at least 100, and roughly one in a thousand will have size at least 1 000 000. At the critical value, the process is poised between extinction and survival.

Another important attribute of the branching process is the average size of a family tree, denoted $\chi(p)$. A calculation shows that

$$\chi(p) = \begin{cases} \frac{1}{1 - 2p} & \text{if } p < \frac{1}{2}, \\ \infty & \text{if } p \geq \frac{1}{2}. \end{cases}$$

In particular, the average family size becomes infinite at the same critical value $p_c = \frac{1}{2}$ above which the probability of an infinite family ceases to be zero. The graph of χ is shown in figure 3. At $p = p_c$, it may seem at first sight contradictory that family trees are always finite (since $\theta(p_c) = 0$) and yet the average family size is infinite (since $\chi(p_c) = \infty$). However, there is no inconsistency, and this combination, which occurs only at the critical point, reflects the slowness of the square-root decay of $P_k(p_c)$.

2.2 Critical Exponents and Universality

Some aspects of the above discussion are specific to twofold branching, and will change for a branching process with higher-order branching. For example, if each organism has not two but m potential offspring, again independently with probability p , then the average number of offspring per organism is mp and the critical probability p_c changes to $1/m$. Also, the formulas written above for the survival probability, for the probability of at least k descendants, and for the average family size must all be modified and will involve the parameter m .

However, the way that $\theta(p)$ goes to zero at the critical point, the way that $P_k(p_c)$ goes to zero as k goes to infinity, and the way that $\chi(p)$ diverges to infinity as p approaches the critical point p_c will all be governed by exponents that are independent of m . To be more specific, they behave in the following manner:

$$\begin{aligned}\theta(p) &\sim C_1(p - p_c)^\beta, & \text{as } p \rightarrow p_c^+, \\ P_k(p_c) &\sim C_2 k^{-1/\delta}, & \text{as } k \rightarrow \infty, \\ \chi(p) &\sim C_3(p_c - p)^{-\gamma}, & \text{as } p \rightarrow p_c^-\end{aligned}$$

Here, the numbers C_1 , C_2 , and C_3 are constants that depend on m . By contrast, the exponents β , δ , and γ take on the same values for every $m \geq 2$. Indeed, those values are $\beta = 1$, $\delta = 2$, and $\gamma = 1$. They are called *critical exponents*, and they are *universal* in the sense that they do not depend on the precise form of the law that governs how the individual organisms reproduce. Related exponents will appear below in other models.

3 Random Graphs

An active research field in discrete mathematics with many applications is the study of objects known as **GRAPHS** [III.34]. These are used to model systems such as the Internet, the World Wide Web, and highway networks. Mathematically, a *graph* is a collection of *vertices* (which might represent computers, Web pages, or cities) joined in pairs by *edges* (physical connections between computers, hyperlinks between Web pages, highways). Graphs are also called *networks*, vertices are also called *nodes* or *sites*, and edges are also called *links* or *bonds*.

3.1 The Basic Model of a Random Graph

A major subarea of graph theory, initiated by Erdős and Rényi in 1960, concerns the properties that a graph typically has when it has been generated randomly. A natural way to do this is to take n vertices and for each

pair to decide randomly (by the toss of a coin, say) whether it should be linked by an edge. More generally, one can choose a number p between 0 and 1 and let p be the probability that any given pair is linked. (This would correspond to using a biased coin to make the decisions.) The properties of random graphs come into their own when n is large, and of particular interest is the fact that there is a phase transition.

3.2 The Phase Transition

If x and y are vertices in a graph, then a *path* from x to y is a sequence of vertices that starts with x and ends with y in such a way that neighboring terms of the sequence are joined by edges. (If the vertices are represented by points and the edges by lines, then a path is a way of getting from x to y by traveling along the lines.) If x and y are joined by a path, then they are said to be *connected*. A *component*, or *connected cluster*, in a graph is what you obtain if you take a vertex together with all the other vertices that are connected to it.

Any graph decomposes naturally into its connected clusters. These will, in general, have different sizes (as measured by the number of vertices), and given a graph it is interesting to know the size of its largest cluster, which we shall denote by N . If we are considering a random graph with n vertices, then the value of N will depend on the multitude of random choices made when the graph was generated, and thus N is itself a random variable. The possible values of N are everything from 1, the value it takes when no edges are present and every cluster consists of a single vertex, to n , when there is just one connected cluster consisting of all the vertices. In particular, $N = 1$ when $p = 0$, and $N = n$ when $p = 1$. At a certain point between these extremes, N undergoes a dramatic jump.

It is possible to guess where the jump might take place, by considering the *degree* of a typical vertex x . This means the number of *neighbors* of x , that is, other vertices that are directly linked to x by a single edge. Each vertex has $n - 1$ potential neighbors, and for each one the probability that it is an actual neighbor is p , so the expected degree of any given vertex is $p(n - 1)$. When p is less than $1/(n - 1)$, each vertex has, on average, less than one neighbor, whereas when p exceeds $1/(n - 1)$, it has, again on average, more than one. This suggests that $p_c = 1/(n - 1)$ will be a critical value, with N being small when p is below p_c , and large when p is above p_c .

This is indeed the case. If we set $p_c = 1/(n-1)$ and write $p = p_c(1 + \varepsilon)$, with ε a fixed number between -1 and $+1$, then $\varepsilon = p(n-1) - 1$. Since $p(n-1)$ is the average degree of each vertex, ε is a measure of how much the average degree differs from 1. Erdős and Rényi showed that, in an appropriate sense, as n goes to infinity,

$$N \sim \begin{cases} 2\varepsilon^{-2} \log n & \text{if } \varepsilon < 0, \\ An^{2/3} & \text{if } \varepsilon = 0, \\ 2\varepsilon n & \text{if } \varepsilon > 0. \end{cases}$$

The A in the above formula is not a constant but a certain random variable that is independent of n (the distribution of which we have not specified here). When $\varepsilon = 0$ and n is large, the formula will tell us, for any $a < b$, the approximate probability that N lies between $an^{2/3}$ and $bn^{2/3}$. To put it another way, A is the *limiting distribution* of the quantity $n^{-2/3}N$ when $\varepsilon = 0$.

There is a marked difference between the behavior of the functions $\log n$, $n^{2/3}$, and n , for large n . The small clusters present for $p < p_c$ correspond to what is called a *subcritical phase*, whereas in the so-called *supercritical phase*, where $p > p_c$, there is a “giant cluster” whose size is of the same order of magnitude as the entire graph (see figure 4).

It is interesting to consider the “evolution” of the random graph, as p is increased from subcritical to supercritical values. (Here one can imagine more and more edges being randomly added to the graph.) A remarkable coalescence occurs, in which many smaller clusters rapidly merge into a giant cluster whose size is proportional to the size of the entire system. The coalescence is thorough, in the sense that in the supercritical phase the giant cluster dominates everything: indeed, the second-largest cluster is known to have asymptotic size only $2\varepsilon^{-2} \log n$, which makes it far smaller than the giant cluster.

3.3 Cluster Size

For branching processes, we defined the quantity $\chi(p)$ to be the average size of the family tree spawned by an individual when the probability of each potential offspring being born was p . By analogy, for the random graph it is natural to take an arbitrary vertex v and define $\chi(p)$ to be the average size of the connected cluster containing v . Since all the vertices play identical roles, $\chi(p)$ is independent of the particular choice of v . If we fix a value of ε , set $p = p_c(1 + \varepsilon)$, and let n tend to infinity, it turns out that the behavior of $\chi(p)$

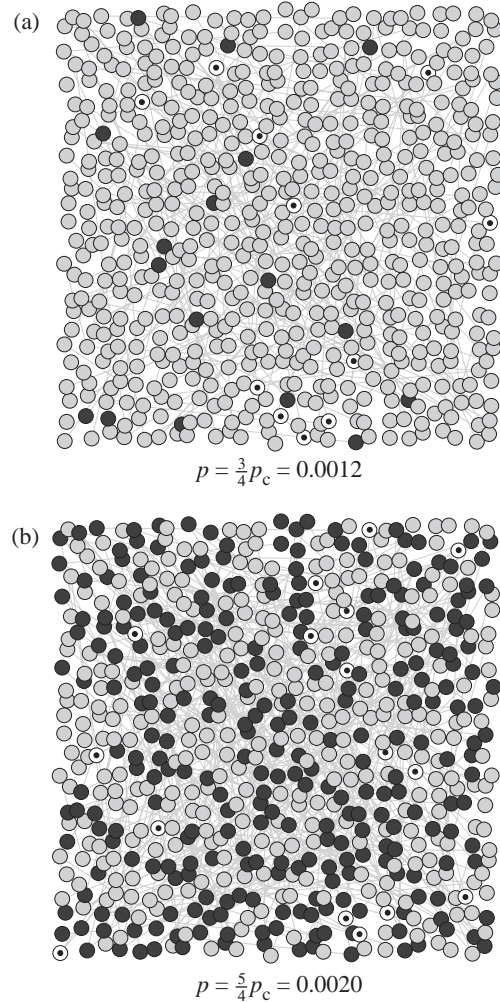


Figure 4 The largest cluster (black) and second largest cluster (dots) in random graphs with 625 vertices. These clusters have sizes (a) 17 and 11 and (b) 284 and 16. The hundreds of edges in the graphs are not clearly shown.

is described by the formula

$$\chi(p) \sim \begin{cases} 1/|\varepsilon| & \text{if } \varepsilon < 0, \\ cn^{1/3} & \text{if } \varepsilon = 0, \\ 4\varepsilon^2 n & \text{if } \varepsilon > 0, \end{cases}$$

where c is a constant. Thus the expected cluster size is independent of n when $\varepsilon < 0$, grows like $n^{1/3}$ when $p = p_c$, and is much larger—indeed, of the same order of magnitude n as the entire system—when $\varepsilon > 0$.

To continue the analogy with branching processes, let $P_k(p)$ denote the probability that the cluster con-

taining the arbitrary vertex v consists of at least k vertices. Again this does not depend on the particular choice of v . In the subcritical phase, when $p = p_c(1 + \varepsilon)$ for some fixed negative value of ε , the probability $P_k(p)$ is essentially independent of n and is exponentially small in k . Thus, large clusters are extremely rare. However, at the critical point $p = p_c$, $P_k(p)$ decays like a multiple of $1/\sqrt{k}$ (for an appropriate range of k). This much slower square-root decay is similar to what happens for branching processes.

3.4 Other Thresholds

It is not only the largest cluster size that jumps. Another quantity that does so is the probability that a random graph is connected, meaning that there is a single connected cluster that contains all the n vertices. For what values of the edge-probability p is this likely? It is known that the property of being connected has a sharp threshold, at $p_{\text{conn}} = (1/n) \log n$, in the following sense. If $p = p_{\text{conn}}(1 + \varepsilon)$ for some fixed negative ε , then the probability that the graph is connected approaches 0 as $n \rightarrow \infty$. If on the other hand ε is positive, then the probability approaches 1. Roughly speaking, if you add edges randomly, then the graph suddenly changes from being almost certainly not connected to almost certainly connected as the proportion of edges present moves from just below p_{conn} to just above it.

There is a wide class of properties with thresholds of this sort. Other examples include the absence of any isolated vertex (a vertex with no incident edge), and the presence of a Hamiltonian cycle (a closed loop that visits every vertex exactly once). Below the threshold, the random graph almost certainly does not have the property, whereas above the threshold it almost certainly does. The transition occurs abruptly.

4 Percolation

The percolation model was introduced by Broadbent and Hammersley in 1957 as a model of fluid flow in a porous medium. The medium contains a network of randomly arranged microscopic pores through which fluid can flow. A d -dimensional medium can be modeled with the help of the infinite d -dimensional lattice \mathbb{Z}^d , which consists of all points x of the form (x_1, \dots, x_d) , where each x_i is an integer. This set can be made into a graph in a natural way if we join each point to the $2d$ points that differ from it by ± 1 in one coordinate and are the same in the others. (So,

for example, in \mathbb{Z}^2 the neighbors of $(2, 3)$ are the four points $(1, 3)$, $(3, 3)$, $(2, 2)$, and $(2, 4)$.) One thinks of the edges as representing all pores potentially present in the medium.

To model the medium itself, one first chooses a *porosity parameter* p , which is a number between 0 and 1. Each edge (or bond) of the above graph is then retained with probability p and deleted with probability $1 - p$, with all choices independent. The retained edges are referred to as “occupied” and the deleted ones as “vacant.” The result is a random subgraph of \mathbb{Z}^d whose edges are the occupied bonds. These model the pores actually present in a macroscopic chunk of the medium.

For fluid to flow through the medium there must be a set of pores connected together on a macroscopic scale. This idea is captured in the model by the existence of an infinite cluster in the random subgraph, that is, a collection of infinitely many points all connected to one another. The basic question is whether or not an infinite cluster exists. If it does, then fluid can flow through the medium on a macroscopic scale, and otherwise it cannot. Thus, when an infinite cluster exists, it is said that “percolation occurs.”

Percolation on the square lattice \mathbb{Z}^2 is depicted in figure 5. Percolation in a three-dimensional physical medium is modeled using \mathbb{Z}^3 . It is instructive, and mathematically interesting, to think how the model’s behavior might change as the dimension d is varied.

For $d = 1$, percolation will not occur unless $p = 1$. The simple observation that leads to this conclusion is the following. Given any particular sequence of m consecutive edges, the probability that they are all occupied is p^m , and if $p < 1$, then this goes to zero as m goes to infinity. The situation is quite different for $d \geq 2$.

4.1 The Phase Transition

For $d \geq 2$, there is a phase transition. Let $\theta(p)$ denote the probability that any given vertex of \mathbb{Z}^d is in an infinite connected cluster. (This probability does not depend on the choice of vertex.) It is known that for $d \geq 2$ there is a critical value p_c , depending on d , such that $\theta(p)$ is zero if $p < p_c$ and positive if $p > p_c$. The exact value of p_c is not known in general, but a special symmetry of the square lattice allows for a proof that $p_c = \frac{1}{2}$ when $d = 2$.

Using the fact that $\theta(p)$ is the probability that *any* particular vertex lies in an infinite cluster, it can be shown that when $\theta(p) > 0$ there must be an infinite

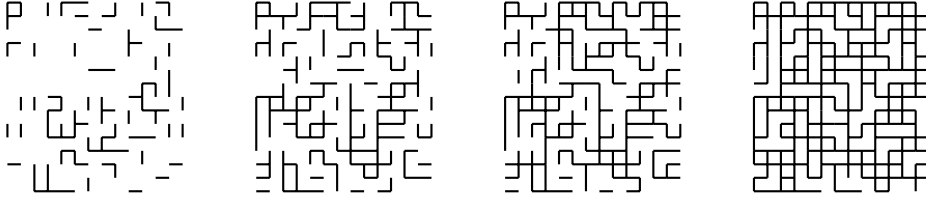


Figure 5 Bond-percolation configurations on a 14×14 piece of the square lattice \mathbb{Z}^2 for $p = 0.25$, $p = 0.45$, $p = 0.55$, $p = 0.75$. The critical value is $p_c = \frac{1}{2}$.

connected cluster somewhere in \mathbb{Z}^d , while when $\theta(p) = 0$ there will not be one. Thus, percolation occurs when $p > p_c$ but not when $p < p_c$, and the system's behavior changes abruptly at the critical value. A deeper argument shows that when $p > p_c$ there must be exactly one infinite cluster; infinite clusters cannot coexist on \mathbb{Z}^d . This is analogous to the situation in the random graph, where one giant cluster dominates when p is above the critical value.

Let $\chi(p)$ denote the average size of the connected cluster containing a given vertex. Certainly $\chi(p)$ is infinite for $p > p_c$, since then there is a positive probability that the given vertex is in an infinite cluster. It is conceivable that $\chi(p)$ could be infinite also for some values of p less than p_c , since infinite expectation is in principle compatible with $\theta(p) = 0$. However, it is a nontrivial and important theorem of the subject that this is not the case: $\chi(p)$ is finite for all $p < p_c$ and diverges to infinity as p approaches p_c from below.

Qualitatively, the graphs of θ and χ have the appearance depicted for the branching process in figures 2 and 3, although the critical value will be less than $\frac{1}{2}$ for $d \geq 3$. There is, however, a caveat. It has been proved that θ is continuous in p except possibly at p_c , and right-continuous for all p . It is widely believed that θ is equal to zero at the critical point, so that θ is continuous for all p and percolation does not occur at the critical point. But proofs that $\theta(p_c) = 0$ are currently known only for $d = 2$, for $d \geq 19$, and for certain related models when $d > 6$. The lack of a general proof is all the more intriguing since it has been proved for all $d \geq 2$ that there is zero probability of an infinite cluster in any half-space when $p = p_c$. This still allows for an infinite cluster with an unnatural spiral behavior, for example, though it is believed that this does not occur.

4.2 Critical Exponents

Assuming that $\theta(p)$ does in fact approach zero as p is decreased to p_c , it is natural to ask in what manner this occurs. Similarly, we can ask in what manner $\chi(p)$ diverges as p increases to p_c . Deep arguments of theoretical physics, and substantial numerical experimentation, have led to the prediction that this, as well as other, behavior is described by certain powers known as *critical exponents*. In particular, it is predicted that there are asymptotic formulas

$$\begin{aligned}\theta(p) &\sim C(p - p_c)^\beta, & \text{as } p \rightarrow p_c^+, \\ \chi(p) &\sim C(p_c - p)^{-\gamma}, & \text{as } p \rightarrow p_c^-\end{aligned}$$

The critical exponents here are the powers β and γ , which depend, in general, on the dimension d . (The letter C is used to denote a constant whose precise value is inessential and may change from line to line.)

When p is less than p_c , large clusters have exponentially small probabilities. For example, in this case the probability $P_k(p)$ that the size of the connected cluster containing any given vertex exceeds k is known to decay exponentially as $k \rightarrow \infty$. At the critical point, this exponential decay is predicted to be replaced by a power-law decay involving a number δ , which is another critical exponent:

$$P_k(p_c) \sim Ck^{-1/\delta} \quad \text{as } k \rightarrow \infty.$$

Also, for $p < p_c$, the probability $\tau_p(x, y)$ that two vertices x and y are in the same connected cluster decays exponentially like $e^{-|x-y|/\xi(p)}$ as the separation between x and y is increased. The number $\xi(p)$ is called the *correlation length*. (Roughly speaking, $\tau_p(x, y)$ starts to become small when the distance between x and y exceeds $\xi(p)$.) The correlation length is known to diverge as p increases to p_c , and the predicted form of this divergence is

$$\xi(p) \sim C(p_c - p)^{-\nu} \quad \text{as } p \rightarrow p_c^-,$$

where ν is a further critical exponent. As before, the decay at the critical point is no longer exponential. It

is predicted that $\tau_{p_c}(x, y)$ decays instead via a power law, traditionally written in the form

$$\tau_{p_c}(x, y) \sim C \frac{1}{|x - y|^{d-2+\eta}}, \quad \text{as } |x - y| \rightarrow \infty,$$

for yet another critical exponent η .

The critical exponents describe large-scale aspects of the phase transition and thus provide information relevant to the macroscopic scale of the physical medium. However, in most cases they have not been rigorously proved to exist. To do so, and to establish their values, is a major open problem in mathematics, one of central importance for percolation theory.

In view of this, it is important to be aware of a prediction from theoretical physics that the exponents are not independent, but are related to each other by what are called *scaling relations*. Three scaling relations are

$$\gamma = (2 - \eta)\nu, \quad \gamma + 2\beta = \beta(\delta + 1), \quad d\nu = \gamma + 2\beta.$$

PUP: I can confirm that repetition of ' $\gamma + 2\beta$ ' is OK.

4.3 Universality

Since the critical exponents describe large-scale behavior, it seems plausible that they might depend only weakly on changes in the fine structure of the model. In fact, it is a further prediction of theoretical physics, one that has been verified by numerical experiments, that the critical exponents are *universal*, in the sense that they depend on the spatial dimension d but on little else.

For example, if the two-dimensional lattice \mathbb{Z}^2 is replaced by another two-dimensional lattice, such as the triangular or the hexagonal lattice, then the values of the critical exponents are believed not to change. Another modification, for general $d \geq 2$, is to replace the standard percolation model with the so-called *spread-out model*. In the spread-out model, the edge set of \mathbb{Z}^d is enriched so that now two vertices are joined whenever they are separated by a distance of L or less, where $L \geq 1$ is a fixed finite parameter, usually taken to be large. Universality suggests that the critical exponents for percolation in the spread-out model do not depend on the parameter L .

The discussion so far falls within the general framework of *bond percolation*, in which it is bonds (edges) that are randomly occupied or vacant. A much-studied variant is *site percolation*, where now it is vertices, or “sites,” that are independently “occupied” with probability p and “vacant” with probability $1 - p$. The connected cluster of a vertex x consists of the vertex x itself together with those occupied vertices that can

be reached by a path that starts at x , travels along edges in the graph, and visits only occupied vertices. For $d \geq 2$, site percolation also experiences a phase transition. Although the critical value for site percolation is different from the critical value for bond percolation, it is a prediction of universality that site and bond percolation on \mathbb{Z}^d have the *same* critical exponents.

These predictions are mathematically very intriguing: the large-scale properties of the phase transition described by critical exponents appear to be insensitive to the fine details of the model, in contrast to features like the value of critical probability p_c , which depends heavily on such details.

At the time of writing, the critical exponents have been proved to exist, and their values rigorously computed, only for certain percolation models in dimensions $d = 2$ and $d > 6$, while a general mathematical understanding of universality remains an elusive goal.

4.4 Percolation in Dimensions $d > 6$

Using a method known as the *lace expansion*, it has been proved that the critical exponents exist, with values

$$\beta = 1, \quad \gamma = 1, \quad \delta = 2, \quad \nu = \frac{1}{2}, \quad \eta = 0,$$

for percolation in the spread-out model when $d > 6$ and L is large enough. The proof makes use of the fact that vertices in the spread-out model have many neighbors. For the more conventional nearest-neighbor model, where bonds have length 1 and there are fewer neighbors per vertex, results of this type have also been obtained, but only in dimensions $d \geq 19$.

The above values of β , γ , and δ are the same as those observed previously for branching processes. A branching process can be regarded as percolation on an infinite tree rather than on \mathbb{Z}^d , and thus percolation in dimensions $d > 6$ behaves like percolation on a tree. This is an extreme example of universality, in which the critical exponents are also independent of the dimension, at least when $d > 6$.

If the above values for the exponents are substituted into the scaling relation $d\nu = \gamma + 2\beta$, the result is $d = 6$. Thus, the scaling relation (called a *hyperscaling* relation because of the presence of the dimension d in the equation) is false for $d > 6$. However, this particular relation is predicted to apply only in dimensions $d \leq 6$. In lower dimensions, the nature of the phase transition is affected by the manner in which critical clusters fit into space, and the nature of the fit is partly described by the hyperscaling relation, in which d appears explicitly.

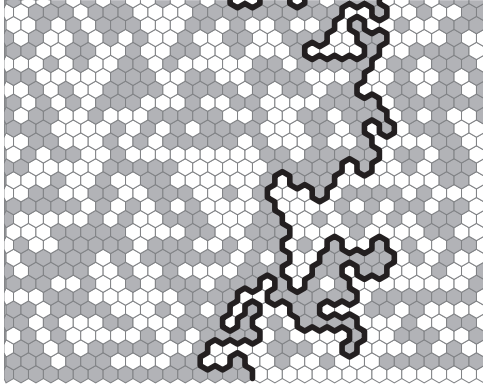


Figure 6 The exploration process.

The critical exponents are predicted to take on different values below $d = 6$. Recent advances have shed much light on the situation for $d = 2$, as we shall see in the next section.

4.5 Percolation in Dimension 2

4.5.1 Critical Exponents and Schramm-Loewner Evolution

For site percolation on the two-dimensional triangular lattice it has been shown, in a major recent achievement, that the critical exponents exist and take the remarkable values

$$\beta = \frac{5}{36}, \quad \gamma = \frac{43}{18}, \quad \delta = \frac{91}{5}, \quad \nu = \frac{4}{3}, \quad \eta = \frac{5}{24}.$$

The scaling relations play an important role in the proof, but an essential additional step requires understanding of a concept known as the *scaling limit*.

To get some idea of what this is, let us look at the so-called *exploration process*, which is depicted in figure 6. In figure 6, hexagons represent vertices of the triangular lattice. Hexagons in the bottom row have been colored gray on the left half and white on the right half. The other hexagons have been chosen to be gray or white independently with probability $\frac{1}{2}$, which is the critical probability for site percolation on the triangular lattice. It is not hard to show that there is a path, also illustrated in figure 6, which starts at the bottom and all along its length is gray to the left and white to the right. The exploration process is this random path, which can be thought of as the gray/white interface. The boundary conditions at the bottom force it to be infinite.

The exploration process provides information about the boundaries separating large critical clusters of dif-

ferent color, and from this it is possible to extract information about critical exponents. It is the macroscopic large-scale structure that is essential, so interest is focused on the exploration process in the limit as the spacing between vertices of the triangular lattice goes to zero. In other words, what does the curve in figure 6 typically look like in the limit as the size of the hexagons shrinks to zero? It is now known that this limit is described by a newly discovered STOCHASTIC PROCESS [IV.24 §1] called the Schramm-Loewner evolution (SLE) with parameter six, or SLE_6 for short. The SLE processes were introduced by Schramm in 2000, and have become a topic of intense current research activity.

This is a major step forward in the understanding of two-dimensional site percolation on the triangular lattice, but much remains to be done. In particular, it is still an unsolved problem to prove universality. There is currently no proof that critical exponents exist for bond percolation on the square lattice \mathbb{Z}^2 , although universality predicts that the critical exponents for the square lattice should also take on the interesting values listed above.

4.5.2 Crossing Probabilities

In order to understand two-dimensional percolation, it is very helpful to understand the probability that there will be a path from one side of a region of the plane to another, especially when the parameter p takes its critical value p_c .

To make this idea precise, fix a simply connected region in the plane (i.e., a region with no holes), and fix two arcs on the boundary of the region. The *crossing probability* (which depends on p) is the probability that there is an occupied path inside the region that joins one arc to the other, or more accurately the limit of this probability as the lattice spacing between vertices is reduced to zero. For $p < p_c$, clusters with diameter much larger than the correlation length $\xi(p)$ (measured by the number of steps in the lattice) are extremely rare. However, to cross the region, a cluster needs to be larger and larger as the lattice spacing goes to zero. It follows that the crossing probability is 0. When $p > p_c$, there is exactly one infinite cluster, from which it can be deduced that if the lattice spacing is very small, then with very high probability there will be a crossing of the region. In the limit, the crossing probability is 1. What if $p = p_c$? There are three remarkable predictions for critical crossing probabilities.

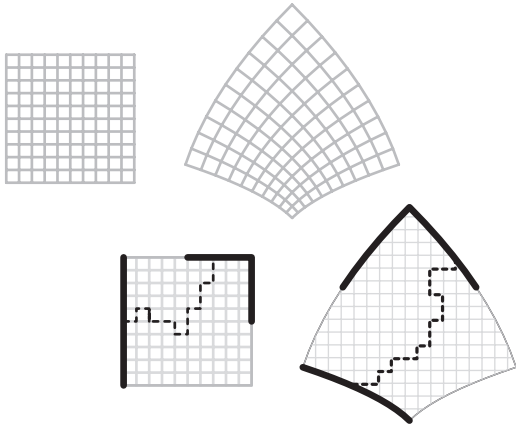


Figure 7 The two regions are related by a conformal transformation, depicted in the upper figures. In the lower figures, the limiting critical crossing probabilities are identical.

The first prediction is that critical crossing probabilities are universal, which is to say that they are the same for all finite-range two-dimensional bond- or site-percolation models. (As always, we are talking about the limiting probabilities as the lattice spacing goes to zero.)

The second prediction is that the critical crossing probabilities are *conformally invariant*. A conformal transformation is a transformation that locally preserves angles, as shown in figure 7. The remarkable RIEMANN MAPPING THEOREM [V.37] states that *any* two simply connected regions that are not the entire plane are related by a conformal transformation. The statement that the critical crossing probability is conformally invariant means that if one region with two specified boundary arcs is mapped to another region by a conformal transformation, then the critical crossing probability between the images of the arcs in the new region is identical to the crossing probability of the original region. (Note that the underlying lattice is *not* transformed; this is what makes the prediction so striking.)

The third prediction is Cardy's explicit formula for critical crossing probabilities. Assuming conformal invariance, it is only necessary to give the formula for one region. For an equilateral triangle, Cardy's formula is particularly simple (see figure 8).

In 2001, in a celebrated achievement, Smirnov studied critical crossing probabilities for site percolation on the triangular lattice. Using the special symmetries of this particular model, Smirnov proved that the lim-

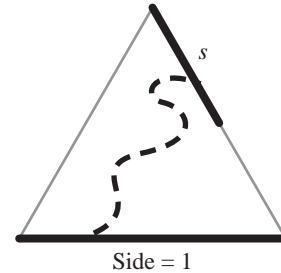


Figure 8 For the equilateral triangle of unit side length, Cardy's formula asserts that the limiting critical crossing probability shown is simply the length s .

iting critical crossing probabilities exist, that they are conformally invariant, and that they obey Cardy's formula. To prove universality of the crossing probabilities remains a tantalizing open problem.

5 The Ising Model

In 1925, Ising published an analysis of a mathematical model of ferromagnetism which now bears his name (although it was in fact Ising's doctoral supervisor Lenz who first defined the model). The Ising model occupies a central position in theoretical physics, and is of considerable mathematical interest.

5.1 Spins, Energy, and Temperature

In the Ising model, a block of iron is regarded as a collection of atoms whose positions are fixed in a crystalline lattice. Each atom has a magnetic "spin," which is assumed for simplicity to point upward or downward. Each possible configuration of spins has an associated energy, and the greater this energy is, the less likely the configuration is to occur.

On the whole, atoms like to have the same spin as their immediate neighbors, and the energy reflects this: it increases according to the number of pairs of neighboring spins that are *not* aligned with each other. If there is an external magnetic field, also assumed to be directed up or down, then there is an additional contribution: atoms like to be aligned with the external field, and the energy is greater the more spins there are that are not aligned with it. Since configurations with higher energy are less likely, spins have a general tendency to align with each other, and also to align with the direction of the external magnetic field. When a larger fraction of spins points up than down, the iron is said to have a positive magnetization.

Although energy considerations favor configurations with many aligned spins, there is a competing effect. As the temperature increases, there are more random thermal fluctuations of the spins, and these diminish the amount of alignment. Whenever there is an external magnetic field, the energy effects predominate and there is at least some magnetization, however high the temperature. However, when the external field is turned off, the magnetization persists only if the temperature is below a certain critical temperature. Above this temperature, the iron will lose its magnetization.

The Ising model is a mathematical model that captures the above picture. The crystalline lattice is modeled by the lattice \mathbb{Z}^d . Vertices of \mathbb{Z}^d represent atomic positions, and the atomic spin at a vertex x is simply modeled by one of the two numbers $+1$ (representing spin up) or -1 (representing spin down). The particular number chosen at x is denoted σ_x , and a collection of choices, one for each x in the lattice, is called a *configuration* of the Ising model. The configuration as a whole is denoted simply as σ . (Formally, a configuration σ is a function from the lattice to the set $\{-1, 1\}$.)

Each configuration σ comes with an associated energy, defined as follows. If there is no external field, the energy of σ consists of the sum, taken over all pairs of neighboring vertices $\langle x, y \rangle$, of the quantity $-\sigma_x \sigma_y$. This quantity is -1 if $\sigma_x = \sigma_y$, and is $+1$ otherwise, so the energy is indeed larger the more nonaligned pairs there are. If there is a nonzero external field, modeled by a real number h , then the energy receives an additional contribution $-h\sigma_x$, which is larger the more spins there are with a different sign from that of h . Thus, in total, the energy $E(\sigma)$ of a spin configuration σ is defined by

$$E(\sigma) = - \sum_{\langle x, y \rangle} \sigma_x \sigma_y - h \sum_x \sigma_x,$$

where the first sum is over neighboring pairs of vertices, the second sum is over vertices, and h is a real number that may be positive, negative, or zero.

The sums defining $E(\sigma)$ actually make sense only when there are finitely many vertices, but one wishes to study the infinite lattice \mathbb{Z}^d . This problem is handled by restricting \mathbb{Z}^d to a large finite subset and later taking an appropriate limit, the so-called *thermodynamic limit*. This is a well-understood process that will not be described here.

Two features remain to be modeled, namely, the manner in which lower-energy configurations are “preferred,” and the manner in which thermal fluctuations

can lessen this preference. Both features are handled simultaneously, as follows. We wish to assign to each configuration a probability that decreases as its energy increases. According to the foundations of statistical mechanics, the right way to do this is to make the probability proportional to the so-called *Boltzmann factor* $e^{-E(\sigma)/T}$, where T is a nonnegative parameter that represents the temperature. Thus, the probability is

$$P(\sigma) = \frac{1}{Z} e^{-E(\sigma)/T},$$

where the normalization constant, or *partition function*, Z , is defined by

$$Z = \sum_{\sigma} e^{-E(\sigma)/T},$$

where the sum is taken over all possible configurations σ (again it is necessary to work first in a finite subset of \mathbb{Z}^d to make this precise). The reason for this choice of Z is that once we divide by it then we have ensured that the probabilities of the configurations add up to one, as they must. With this definition, the desired preference for low energy is achieved, since the probability of a given configuration is smaller when the energy of the configuration is larger. As for the effect of the temperature, note that when T is very large, all the numbers $e^{-E(\sigma)/T}$ are close to 1, so all probabilities are roughly equal. In general, as the temperature increases the probabilities of the various configurations become more similar, and this models the effect of random thermal fluctuations.

There is more to the story than energy, however. The Boltzmann factor makes any individual low-energy configuration much more likely than any individual high-energy configuration. However, the low-energy configurations have a high degree of alignment, so there are far fewer of them than there are of the more randomly arranged high-energy configurations. It is not obvious which of these two competing considerations will predominate, and in fact the answer depends on the value of the temperature T in a very interesting way.

5.2 The Phase Transition

For the Ising model with external field h and temperature T , let us choose a configuration randomly with the probabilities defined above. The *magnetization* $M(h, T)$ is defined to be the expected value of the spin σ_x at a given vertex x . Because of the symmetry of the lattice \mathbb{Z}^d , this does not depend on the particular vertex chosen. Accordingly, if the magnetization $M(h, T)$ is positive, then spins have an overall tendency

to be aligned in the positive direction, and the system is magnetized.

The symmetry between up and down implies that $M(-h, T) = -M(h, T)$ (i.e., reversing the external field reverses the magnetization) for all h and T . In particular, when $h = 0$, the magnetization must be zero. On the other hand, if there is a nonzero external field h , then configurations with spins that are aligned with h are overwhelmingly more likely (because their energy is lower), and the magnetization satisfies

$$M(h, T) \begin{cases} < 0 & \text{if } h < 0, \\ = 0 & \text{if } h = 0, \\ > 0 & \text{if } h > 0. \end{cases}$$

What happens if the external field is initially positive and then is reduced to zero? In particular, is the *spontaneous magnetization*, defined by

$$M_+(T) = \lim_{h \rightarrow 0^+} M(h, T),$$

positive or zero? If $M_+(T)$ is positive, then the magnetization persists after the external field is turned off. In this case there will be a discontinuity in the graph of M versus h at $h = 0$.

Whether or not this happens depends on the temperature T . In the limit as T is reduced to zero, a small difference in the energies of two configurations results in an enormous difference in their probabilities. When $h > 0$ and the temperature is reduced to zero, only the minimal energy configuration, in which all spins are $+1$, has any chance of occurring. This is the case no matter how small the external field becomes, so $M_+(0) = 1$. On the other hand, in the limit of infinitely high temperature, all configurations become equally likely and the spontaneous magnetization is equal to zero.

For dimensions $d \geq 2$, the behavior of $M_+(T)$ when T lies between these two extremes is quite surprising. In particular, it is not differentiable everywhere: there is a critical temperature T_c , depending on the dimension, such that the spontaneous magnetization is strictly positive for $T < T_c$ and zero for $T > T_c$, and it is at $T = T_c$ that differentiability fails. Schematic graphs of the magnetization versus h and the spontaneous magnetization versus T are shown in figure 9. What happens at the critical temperature itself is delicate. In all dimensions except $d = 3$ it has been proved that there is no spontaneous magnetization at the critical temperature, which is to say that $M_+(T_c) = 0$. It is believed that this is true when $d = 3$ as well, but it remains an open problem to prove it.

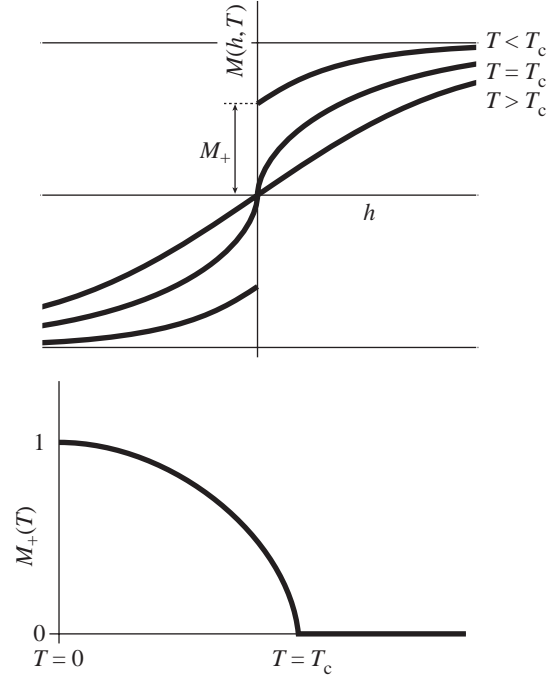


Figure 9 Magnetization versus external field, and spontaneous magnetization versus temperature.

5.3 Critical Exponents

The phase transition for the Ising model is again described by critical exponents. The critical exponent β , given by

$$M_+(T) \sim C(T_c - T)^\beta, \quad \text{as } T \rightarrow T_c^-,$$

indicates how the spontaneous magnetization disappears as the temperature increases toward the critical temperature T_c . For $T > T_c$, the *magnetic susceptibility*, denoted $\chi(T)$, is defined to be the rate of change of $M(h, T)$ with respect to h , at $h = 0$. This partial derivative in h diverges as T approaches T_c from above, and the exponent γ is defined by

$$\chi(T) \sim C(T - T_c)^{-\gamma}, \quad \text{as } T \rightarrow T_c^+.$$

Finally, δ describes the manner in which the magnetization goes to zero as the external field is reduced to zero at the critical temperature. That is,

$$M(h, T_c) \sim Ch^{1/\delta}, \quad \text{as } h \rightarrow 0^+.$$

These critical exponents, like those for percolation, are predicted to be universal and to obey various scaling relations. They are now understood mathematically in all dimensions except $d = 3$.

5.4 Exact Solution for $d = 2$

In 1944, Onsager published a famous paper in which he gave an exact solution of the two-dimensional Ising model. His remarkable computation is a landmark in the development of the theory of critical phenomena. With the exact solution as a starting point, critical exponents could be calculated. As with two-dimensional percolation, the exponents take interesting values:

$$\beta = \frac{1}{8}, \quad \gamma = \frac{7}{4}, \quad \delta = 15.$$

5.5 Mean-Field Theory for $d \geq 4$

Two modifications of the Ising model are relatively easy to analyze. One is to formulate the model on the infinite binary tree, rather than on the integer lattice \mathbb{Z}^d . Another is to formulate the Ising model on the so-called “complete graph,” which is the graph consisting of n vertices with an edge joining every pair of vertices, and then take the limit as n goes to infinity. In the latter, known as the *Curie-Weiss model*, each spin interacts equally with all the other spins, or, put another way, each spin feels the *mean field* of all the other spins. In each of these modifications, the critical exponents take on the so-called mean-field values

$$\beta = \frac{1}{2}, \quad \gamma = 1, \quad \delta = 3.$$

Ingenious methods have been used to prove that the Ising model on \mathbb{Z}^d has these same critical exponents in dimensions $d \geq 4$, although in dimension 4 there remain unresolved issues concerning logarithmic corrections to the asymptotic formulas.

6 The Random-Cluster Model

The percolation and Ising models appear to be quite different. A percolation configuration consists of a random subgraph of a given graph (usually a lattice as in the examples earlier), with edges included independently with probability p . A configuration of the Ising model consists of an assignment of values ± 1 to spins at the vertices of a graph (again usually a lattice), with these spins influenced by energy and temperature.

In spite of these differences, in around 1970 Fortuin and Kasteleyn had the insight to observe that the two models are in fact closely related to each other, as members of a larger family of models known as the random-cluster model. The random-cluster model also includes a natural extension of the Ising model known as the *Potts model*.

In the Potts model, spins at the vertices of a given graph G may take on any one of q different values,

where q is an integer greater than or equal to 2. When $q = 2$ there are two possible spin values and the model is equivalent to the Ising model. For general q , it is convenient to label the possible spin values as $1, 2, \dots, q$. As before, a configuration of spins has an associated energy that is smaller when more spins are aligned. The energy associated with an edge is -1 if the spins at the vertices joined by the edge are identical, and 0 otherwise. The total energy $E(\sigma)$ of a spin configuration σ , assuming no external field, is the sum of the energies associated with all edges. The probability of a particular spin configuration σ is again taken to be proportional to a Boltzmann factor, namely

$$P(\sigma) = \frac{1}{Z} e^{-E(\sigma)/T},$$

where the partition function Z is once again there to ensure that the probabilities add up to 1.

Fortuin and Kasteleyn noticed that the partition function of the Potts model on a finite graph G can be recast as

$$\sum_{S \subset G} p^{|S|} (1-p)^{|G \setminus S|} q^{n(S)}.$$

In this formula, the sum is over all subgraphs S that can be obtained by deleting edges from G , $|S|$ is the number of edges in S , $|G \setminus S|$ is the number of edges deleted from G to obtain S , $n(S)$ is the number of distinct connected clusters of S , and p is related to the temperature by

$$p = 1 - e^{-1/T}.$$

The restriction that q be an integer greater than or equal to 2 is essential for the definition of the Potts model, but the above sum makes good sense for any positive real value of q .

The *random-cluster model* has the above sum as its partition function. Given any real number $q > 0$, a configuration of the random-cluster model is a set S of occupied edges of the graph G , exactly like a configuration of bond percolation. However, in the random-cluster model we do not simply associate p with each occupied edge and $1-p$ with each vacant edge. Instead, the probability associated with a configuration is proportional to $p^{|S|} (1-p)^{|G \setminus S|} q^{n(S)}$. In particular, for the choice $q = 1$, the random-cluster model is the same as bond percolation. Thus the random-cluster model provides a one-parameter family of models, indexed by q , which corresponds to percolation for $q = 1$, to the Ising model for $q = 2$, and to the Potts model for integer $q \geq 2$. The random-cluster model has a phase transition for general $q \geq 1$, and provides a unified setting and a rich family of examples.

7 Conclusion

The science of critical phenomena and phase transitions is a source of fascinating mathematical problems of real physical significance. Percolation is a central mathematical model in the subject. Often formulated on \mathbb{Z}^d , it can also be defined instead on a tree or on the complete graph, as a result of which it encompasses branching processes and the random graph. The Ising model is a fundamental model of the ferromagnetic phase transition. At first sight unrelated to percolation, it is in fact closely connected within the wider setting of the random-cluster model. The latter provides a unified framework and a powerful geometric representation for the Ising and Potts models.

Part of the fascination of these models is due to the prediction from theoretical physics that large-scale features near the critical point are universal. However, proofs often rely on specific details of a model, even when universality predicts that these details should not be essential to the results. For example, the understanding of critical crossing probabilities and the calculation of critical exponents has been carried out for site percolation on the triangular lattice, but not for bond percolation on \mathbb{Z}^2 . Although the progress for the triangular lattice is a triumph of the theory, it is not the last word. Universality remains a guiding principle but it is not yet a general theorem.

In the physically most interesting case of dimension 3, a very basic feature of percolation and the Ising model is not understood at all: it has not yet been proved that there is no percolation at the critical point and that the spontaneous magnetization is zero.

Much has been accomplished but much remains to be done, and it seems clear that further investigation of models of critical phenomena will lead to highly important mathematical discoveries.

Acknowledgments. The figures were produced by Bill Caselman, Department of Mathematics, University of British Columbia, and Graphics Editor of *Notices of the American Mathematical Society*.

Further Reading

- Grimmett, G. R. 1999. *Percolation*, 2nd edn. New York: Springer.
- . 2004. The random-cluster model. In *Probability on Discrete Structures*, edited by H. Kesten, pp. 73–124. New York: Springer.
- Janson, S., T. Łuczak, and A. Ruciński. 2000. *Random Graphs*. New York: John Wiley.

Thompson, C. J. 1988. *Classical Equilibrium Statistical Mechanics*. Oxford: Oxford University Press.

Werner, W. 2004. Random planar curves and Schramm-Loewner evolutions. In *Lectures on Probability Theory and Statistics. École d'Été de Probabilités de Saint-Flour XXXII-2002*, edited by J. Picard. Lecture Notes in Mathematics, volume 1840. New York: Springer.

IV.26 High-Dimensional Geometry and Its Probabilistic Analogues

Keith Ball

1 Introduction

If you have ever watched a child blowing soap bubbles, then you cannot have failed to notice that the bubbles are, at least as far as the human eye can tell, perfectly spherical. From a mathematical perspective, the reason for this is simple. The surface tension in the soap solution causes each bubble to make its area as small as possible, subject to the constraint that it encloses a fixed amount of air (and cannot compress the air too much). The sphere is the surface of smallest area that encloses a given volume.

As a mathematical principle, this seems to have been recognized by the ancient Greeks, although fully rigorous demonstrations did not appear until the end of the nineteenth century. This and similar statements are known as “isoperimetric principles.”¹

The two-dimensional form of the problem asks: what is the shortest curve that encloses a given area? The answer, as we might expect by analogy with the three-dimensional case, is a circle. Thus, by minimizing the length of the curve we force it to have a great deal of symmetry: the curve should be equally curved everywhere along its length. In three or more dimensions, many different kinds of CURVATURE [III.80] are used in different contexts. One, known as *mean curvature*, is the appropriate one for area-minimization problems.

The sphere has the same mean curvature at every point, but then it is pretty clear from its symmetry that the sphere would have the same curvature at every point whatever measure of curvature we used. More illustrative examples are provided by the soap films (much more varied than simple bubbles) that are a popular feature of recreational mathematics lectures: figure 1 shows such a soap film stretched across a wire

1. The prefix “iso” means “equal.” The name “equal perimeter” refers to the two-dimensional formulation: if a disk and another region have equal perimeter, then the area of the other region cannot be larger than that of the disk.



Figure 1 A soap film has minimum area.

frame. The film adopts the shape that minimizes its area, subject to the constraint that it is bounded by the wire frame. One can show that the minimal surface (the exact mathematical solution to the minimization problem) has constant mean curvature: its mean curvature is the same at every point.

Isoperimetric principles turn up all over mathematics: in the study of partial differential equations, the calculus of variations, harmonic analysis, computational algorithms, probability theory, and almost every branch of geometry. The aim of the first part of this article is to describe a branch of mathematics, high-dimensional geometry, whose starting point is the fundamental isoperimetric principle: that the sphere is the surface of least area that encloses a given volume. The most remarkable feature of high-dimensional geometry is its intimate connection to the theory of probability: geometric objects in high-dimensional space exhibit many of the characteristic properties of random distributions. The aim of the second part of this article is to outline the links between the geometry and probability.

2 High-Dimensional Spaces

So far we have discussed only two- and three-dimensional geometry. Higher-dimensional spaces seem to be impossible for humans to visualize but it is easy to provide a mathematical description by extending the usual description of three-dimensional space in terms of Cartesian coordinates. In three dimensions, a point (x, y, z) is given by three coordinates; in n -dimensional space, the points are n -tuples (x_1, x_2, \dots, x_n) . As in two and three dimensions, the points are related to one another in that we can add two of them together to produce a third, by simply adding corresponding coordinates:

$$(2, 3, \dots, 7) + (1, 5, \dots, 2) = (3, 8, \dots, 9).$$

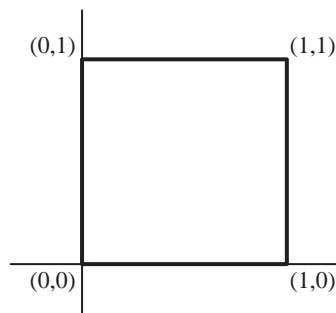


Figure 2 The unit square.

By relating points to one another, addition gives the space some structure or “shape.” The space is not just a jumble of unrelated points.

To describe the shape of the space completely, we also need to specify the distance between any two points. In two dimensions, the distance of a point (x, y) from the origin is $\sqrt{x^2 + y^2}$ by the Pythagorean theorem (and the fact that the axes are perpendicular). Similarly, the distance between two points (u, v) and (x, y) is

$$\sqrt{(x - u)^2 + (y - v)^2}.$$

In n dimensions we define the distance between points (u_1, u_2, \dots, u_n) and (x_1, x_2, \dots, x_n) to be

$$\sqrt{(x_1 - u_1)^2 + (x_2 - u_2)^2 + \dots + (x_n - u_n)^2}.$$

Volume is defined in n -dimensional space roughly as follows. We start by defining a cube in n dimensions. The two- and three-dimensional cases, the square and the usual three-dimensional cube, are very familiar. The set of all points in the xy -plane whose coordinates are between 0 and 1 is a square of side 1 unit (as shown in figure 2), and, similarly, the set of all points (x, y, z) for which x, y , and z are all between 0 and 1 is a unit cube. In n -dimensional space the analogous cube consists of those points whose coordinates are all between 0 and 1. We stipulate that the unit cube has volume 1. Now, if we double the size of a plane figure, its area increases by a factor of 4. If we double a three-dimensional body, its volume increases by a factor of 8. In n -dimensional space, the volume scales as the n th power of size: so a cube of side t has volume t^n . To find the volume of a more general set we try to approximate it by covering it with little cubes whose total volume is as small as possible. The volume of the set is calculated as a limit of these approximate volumes.

Whatever the dimension, a special geometric role is played by the *unit sphere*: that is, the surface consist-

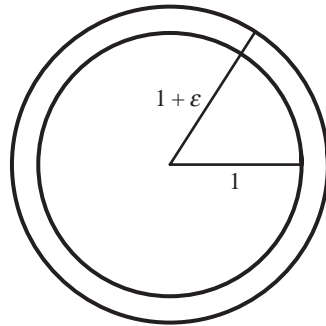


Figure 3 An inflated ball.

ing of all points that are a distance of 1 unit from a fixed point, the center. As one might expect, the corresponding solid sphere, or *unit ball*, consisting of all points enclosed by the unit sphere, also plays a special role. There is a simple relationship between the (n -dimensional) volume of the unit ball and the $(n - 1)$ -dimensional “area” of the sphere. If we let v_n denote the volume of the unit ball in n dimensions, then the surface area is nv_n . One way to see this is to imagine enlarging the unit ball by a factor slightly greater than 1, say $1 + \varepsilon$. This is pictured in figure 3. The enlarged ball has volume $(1 + \varepsilon)^n v_n$ and so the volume of the shell between the two spheres is $((1 + \varepsilon)^n - 1)v_n$. Since the shell has thickness ε , this volume is approximately the surface area multiplied by ε . So the surface area is approximately

$$\frac{(1 + \varepsilon)^n - 1}{\varepsilon} v_n.$$

By taking the limit as ε approaches 0 we obtain the surface area exactly:

$$\lim_{\varepsilon \rightarrow 0} \frac{(1 + \varepsilon)^n - 1}{\varepsilon} v_n.$$

One can check that this limit is nv_n either by expanding the power $(1 + \varepsilon)^n$ or by observing that the expression is the formula for a derivative.

So far we have discussed bodies in n -dimensional space without being too precise about what kind of sets we are considering. Many of the statements in this article hold true for quite general sets. But a special role is played in high-dimensional geometry by convex sets (a set is convex if it contains the entire line segment joining any two of its points). Balls and cubes are both examples of convex sets. The next section describes a fundamental principle which holds for very general sets but which is intrinsically linked to the notion of convexity.

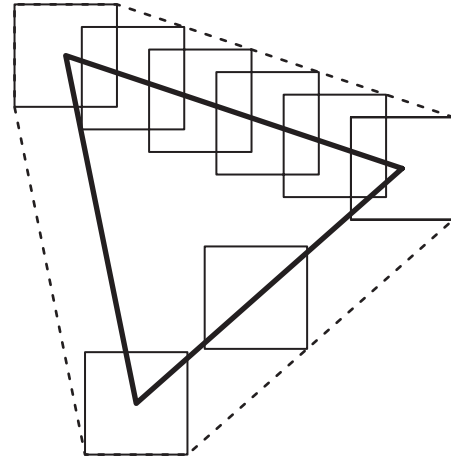


Figure 4 Adding two sets.

3 The Brunn–Minkowski Inequality

The two-dimensional isoperimetric principle was essentially proved in 1841 by Steiner, although there was a technical gap in the argument which was filled later. The general (n -dimensional) case was completed by the end of the nineteenth century. A couple of decades later a different approach to the principle, with far-reaching consequences, was found by HERMANN MINKOWSKI [VI.64]—an approach which was inspired by an idea of Hermann Brunn.

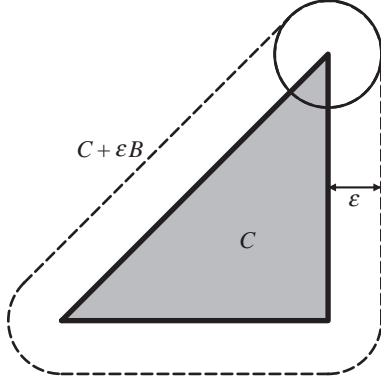
T&T note: check CR style later.

Minkowski considered the following way to add together two *sets* in n -dimensional space. If C and D are sets, then the sum $C + D$ consists of all points which can be obtained by adding a point of C to a point of D . Figure 4 shows an example in which C is an equilateral triangle and D is a square centered at the origin. We place a copy of the square at each point of the triangle (some of these are illustrated) and the set $C + D$ consists of all points that are included in all these squares. The outline of $C + D$ is shown dashed.

The Brunn–Minkowski inequality relates the volume of the sum of two sets to the volumes of the sets themselves. It states that (as long as the two sets C and D are not empty)

$$\text{vol}(C + D)^{1/n} \geq \text{vol}(C)^{1/n} + \text{vol}(D)^{1/n}. \quad (1)$$

The inequality looks a bit technical, if only because the volumes appearing in the inequality are raised to the power $1/n$. However, this fact is crucial. If each of C and D is a unit cube (with their edges aligned the same way), then the sum $C + D$ is a cube of side 2: a cube twice as large. Each of C and D has volume 1 while the

Figure 5 An ε -enlargement.

volume of $C + D$ is 2^n . So, in this case, $\text{vol}(C+D)^{1/n} = 2$ and each of $\text{vol}(C)^{1/n}$ and $\text{vol}(D)^{1/n}$ is equal to 1: the inequality (1) holds *with equality*. Similarly, whenever C and D are copies of one another, the Brunn-Minkowski inequality holds with equality. If we omitted the exponents $1/n$, the statement would still be true; in the case of two cubes, it is certainly true that $2^n \geq 1 + 1$. But the statement would be extremely weak: it would give us almost no useful information.

The importance of the Brunn-Minkowski inequality stems from the fact that it is the most fundamental principle relating volume to the operation of addition, which is the operation that gives space its structure. At the start of this section it was explained that Minkowski's formulation of Brunn's idea provided a new approach to the isoperimetric principle. Let us see why.

Let C be a COMPACT SET [III.9] in \mathbb{R}^n whose volume is equal to that of the unit ball B . We want to show that the surface area of C is at least $n \text{vol}(B)$ since this is the surface area of the ball. We consider what happens to C if we add a small ball to it. An example (a right-angled triangle) is shown in figure 5: the dashed curve outlines the enlarged set we obtain by adding to C a copy of the ball B scaled by a small factor ε . This looks rather like figure 3 above but here we do not expand the original set, we add a ball. Just as before, the difference between $C + \varepsilon B$ and C is a shell around C of width ε , so we can express the surface area as a limit as ε approaches 0:

$$\lim_{\varepsilon \rightarrow 0} \frac{\text{vol}(C + \varepsilon B) - \text{vol}(C)}{\varepsilon}.$$

Now the Brunn-Minkowski inequality tells us that

$$\text{vol}(C + \varepsilon B)^{1/n} \geq \text{vol}(C)^{1/n} + \text{vol}(\varepsilon B)^{1/n}.$$

The right-hand side of this inequality is

$$\text{vol}(C)^{1/n} + \varepsilon \text{vol}(B)^{1/n} = (1 + \varepsilon) \text{vol}(B)^{1/n}$$

because $\text{vol}(\varepsilon B) = \varepsilon^n \text{vol}(B)$ and $\text{vol}(C) = \text{vol}(B)$. So the surface area is at least

$$\begin{aligned} \lim_{\varepsilon \rightarrow 0} \frac{(1 + \varepsilon)^n \text{vol}(B) - \text{vol}(C)}{\varepsilon} \\ = \lim_{\varepsilon \rightarrow 0} \frac{(1 + \varepsilon)^n \text{vol}(B) - \text{vol}(B)}{\varepsilon}. \end{aligned}$$

Again as in section 2, this limit is $n \text{vol}(B)$ and we conclude that the surface of C has at least this area.

Over the years, many different proofs of the Brunn-Minkowski inequality have been found, and most of the methods have other important applications. To finish this section we shall describe a modified version of the Brunn-Minkowski inequality that is often easier to use than (1). If we replace the set $C + D$ by a scaled copy half as large, $\frac{1}{2}(C + D)$, then its volume is scaled by $1/2^n$ and the n th root of this volume is scaled by $\frac{1}{2}$. Therefore, the inequality can be rewritten

$$\text{vol}(\tfrac{1}{2}(C + D))^{1/n} \geq \tfrac{1}{2} \text{vol}(C)^{1/n} + \tfrac{1}{2} \text{vol}(D)^{1/n}.$$

Because of the simple inequality $\frac{1}{2}x + \frac{1}{2}y \geq \sqrt{xy}$ for positive numbers, the right-hand side of this inequality is at least $\sqrt{\text{vol}(C)^{1/n} \text{vol}(D)^{1/n}}$. It follows that

$$\text{vol}(\tfrac{1}{2}(C + D))^{1/n} \geq \sqrt{\text{vol}(C)^{1/n} \text{vol}(D)^{1/n}}$$

and hence that

$$\text{vol}(\tfrac{1}{2}(C + D)) \geq \sqrt{\text{vol}(C) \text{vol}(D)}. \quad (2)$$

We shall elucidate a striking consequence of this inequality in the next section.

The Brunn-Minkowski inequality holds true for very general sets in n -dimensional space, but for convex sets it is the beginning of a surprising theory that was initiated by Minkowski and developed in a remarkable way by Aleksandrov, Fenchel, and Blaschke, among others: the theory of so-called mixed volumes. In the 1970s Khovanskii and Teissier (using a discovery of D. Bernstein) found an astonishing connection between the theory of mixed volumes and the Hodge index theorem in algebraic geometry.

4 Deviation in Geometry

Isoperimetric principles state that if a set is reasonably large, then it has a large surface or boundary. The Brunn-Minkowski inequality (and especially the argument we used to deduce the isoperimetric principle) expands upon this statement by showing that if we start with a reasonably large set and extend it (by adding a small ball), then the volume of the new set is quite a lot bigger than that of the original. During the 1930s Paul Lévy realized that in certain situations,

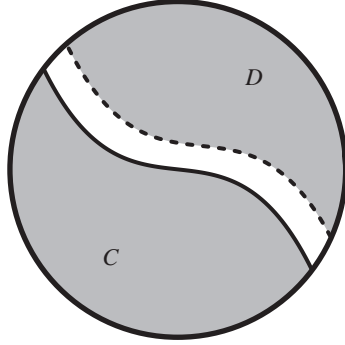


Figure 6 Expanding half a ball.

this fact can have very striking consequences. To get an idea of how this works suppose that we have a compact set C inside the unit ball, whose volume is half that of the ball; for example, C might be the set pictured in figure 6.

Now extend the set C by including all points of the ball that are within distance ε of C , much as we did when deducing the isoperimetric inequality (the dashed curve in figure 6 shows the boundary of the extended set). Let D denote the remainder of the ball (also illustrated). Then if c is a point in C and d is a point in D , we are guaranteed that c and d are separated by a distance of at least ε . A simple two-dimensional argument, pictured in figure 7, shows that in this case the midpoint $\frac{1}{2}(c + d)$ cannot be too near the surface of the ball. In fact, its distance from the center is no more than $1 - \frac{1}{8}\varepsilon^2$. So the set $\frac{1}{2}(C + D)$ lies inside the ball of radius $1 - \frac{1}{8}\varepsilon^2$, whose volume is $(1 - \frac{1}{8}\varepsilon^2)^n$ times the volume of the ball v_n . The crucial point is that if the exponent n is large and ε is not too small, the factor $(1 - \frac{1}{8}\varepsilon^2)^n$ is extremely small: in a space of high dimension, a ball of slightly smaller radius has very much smaller volume. In order to make use of this we apply inequality (2), which states that the volume of $\frac{1}{2}(C + D)$ is at least $\sqrt{\text{vol}(C) \text{vol}(D)}$. Therefore,

$$\sqrt{\text{vol}(C) \text{vol}(D)} \leq (1 - \frac{1}{8}\varepsilon^2)^n v_n$$

or, equivalently,

$$\text{vol}(C) \text{vol}(D) \leq (1 - \frac{1}{8}\varepsilon^2)^{2n} v_n^2.$$

Since the volume of C is $\frac{1}{2}v_n$, we deduce that

$$\text{vol}(D) \leq 2(1 - \frac{1}{8}\varepsilon^2)^{2n} v_n.$$

It is convenient to replace the factor $(1 - \frac{1}{8}\varepsilon^2)^{2n}$ by a (pretty accurate) approximation $e^{-n\varepsilon^2/4}$, which is slightly easier to understand. We can then conclude

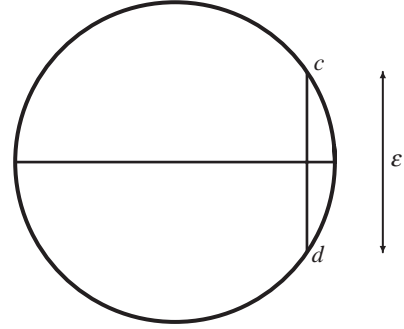


Figure 7 A two-dimensional argument.

that the volume $\text{vol}(D)$ of the residual set D satisfies the inequality

$$\text{vol}(D) \leq 2e^{-n\varepsilon^2/4} v_n. \quad (3)$$

If the dimension n is large, then the exponential factor $e^{-n\varepsilon^2/4}$ is very small, as long as ε is a bit bigger than $1/\sqrt{n}$. What this means is that only a small fraction of the ball lies in the residual set D . All but a small fraction of the ball lies close to C , even though *some* points in the ball may lie much farther from C . Thus, if we start with a set (any set) that occupies half the ball and extend it a little bit, we swallow up almost the entire ball. With a little more sophistication, the same argument can be used to show that the surface of the ball, the sphere, has exactly the same property. If a set C occupies half the sphere, then almost all of the sphere is close to that set.

This counterintuitive effect turns out to be characteristic of high-dimensional geometry. During the 1980s a startling probabilistic picture of high-dimensional space was developed from Lévy's basic idea. This picture will be sketched in the next section.

One can see why the high-dimensional effect has a probabilistic aspect if one thinks about it in a slightly different way. To begin with, let us ask ourselves a basic question: what does it mean to choose a random number between 0 and 1? It could mean many things but if we want to specify one particular meaning, then our job is to decide what the chance is that the random number will fall into each possible range $a \leq x \leq b$: what is the chance that it lies between 0.12 and 0.47, for example? For most people, the obvious answer is 0.35, the difference between 0.47 and 0.12. The probability that our random number lands in the interval $a \leq x \leq b$ will just be $b - a$, the length of that interval. This way of choosing a random number is called *uniform*. Equal-

sized parts of the range between 0 and 1 are equally likely to be selected.

Just as we can use length to describe what is meant by a random number, we can use the volume measure in n -dimensional space to say what it means to select a random point of the n -dimensional ball. We have to decide what the chance is that our random point falls into each subregion of the ball. The most natural choice is to say that it is equal to the volume of that subregion divided by the volume of the entire ball, that is, the proportion of the ball occupied by the subregion. With this choice of random point, it is possible to reformulate the high-dimensional effect in the following way. If we choose a subset C of the ball which has a $\frac{1}{2}$ chance of being hit by our random point, then the chance that our random point lies more than ε away from C is no more than $2e^{-n\varepsilon^2/4}$.

To finish this section it will be useful to rephrase the geometric deviation principle as a statement about functions rather than sets. We know that if C is a set occupying half the sphere, then almost the entire sphere is within a small distance of C . Now suppose that f is a function defined on the sphere: f assigns a real number to each point of the sphere. Assume that f cannot change too rapidly as you move around the sphere: for example, that the values $f(x)$ and $f(y)$ at two points x and y cannot differ by more than the distance between x and y . Let M be the *median* value of f , meaning that f is at most M on half the sphere and at least M on the other half. Then it follows from the deviation principle that f must be almost equal to M on all but a small fraction of the sphere. The reason is that almost all of the sphere is close to the half where f is *below* M ; so f cannot be much *more* than M except on a small set. On the other hand, almost all of the sphere is close to the half where f is *at least* M ; so f cannot be much *less* than M except on a small set.

Thus, the geometric deviation principle says that if a function on the sphere does not vary too fast, then it must be almost constant on almost the entire sphere (even though there may be some points where it is very far from this constant value).

5 High-Dimensional Geometry

It was mentioned at the end of section 3 that convex sets have a special significance in Minkowski's theory relating volume to the additive structure of space. They also occur naturally in a large number of applications: in linear programming and partial differential

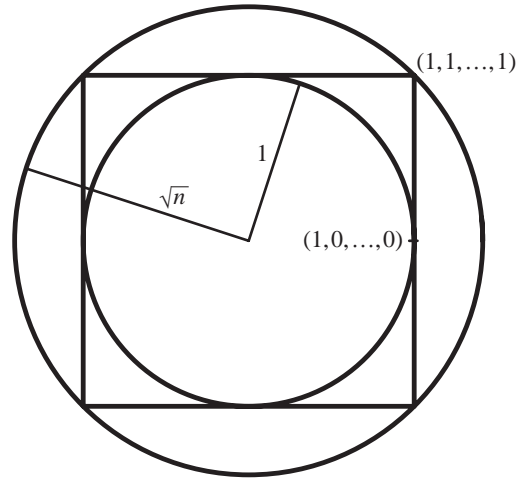


Figure 8 A ball in a box in a ball.

equations, for example. Although convexity is a fairly restrictive condition for a body to satisfy, it is not hard to convince oneself that convex sets exhibit considerable variety and that this variety seems to increase with the dimension. The simplest convex sets after the balls are cubes. If the dimension is large, the surface of a cube looks very unlike the sphere. Let us consider, not a unit cube, but a cube of side 2 whose center is the origin. The corners of the cube are points like $(1, 1, \dots, 1)$ or $(1, -1, -1, \dots, 1)$, whose coordinates are all equal to 1 or -1 , while the center of each *face* is a point like $(1, 0, 0, \dots, 0)$ which has just one coordinate equal to 1 or -1 . The corners are at a distance \sqrt{n} from the center of the cube, while the centers of the faces are at distance 1 from the origin. Thus, the largest sphere that can be fitted inside the cube has radius 1, while the smallest sphere that encloses the cube has radius \sqrt{n} (this is illustrated in figure 8).

When the dimension n is large, this ratio of \sqrt{n} is also large. As one might expect, this gap between the ball and the cube is able to accommodate a wide variety of different convex shapes. Nevertheless, the probabilistic view of high-dimensional geometry has led to an understanding that, for many purposes, this enormous variety is an illusion: that in certain well-defined senses, all convex bodies behave like balls.

Probably the first discovery that pointed strongly in this direction was made by Dvoretzky in the late 1960s. DVORETZKY'S THEOREM [V.10] says that every high-dimensional convex body has slices that are almost spherical. More precisely, if you specify a dimension (say ten) and a degree of accuracy, then for any suffi-

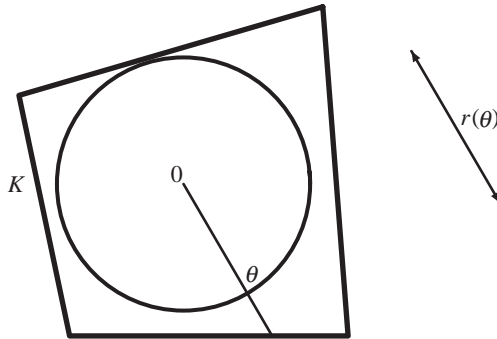


Figure 9 The directional radius.

ciently large dimension n , every n -dimensional convex body has a ten-dimensional slice that is indistinguishable from a ten-dimensional sphere, up to the specified accuracy.

The proof of Dvoretzky's theorem that is conceptually simplest depends upon the deviation principle described in the last section and was found by Milman a few years after Dvoretzky's theorem appeared. The idea is roughly this. Consider a convex body K in n dimensions that contains the unit ball. For each point θ on the sphere, imagine the line segment starting at the origin, passing through the sphere at θ , and extending out to the surface of K (see figure 9). Think of the length of this line as the "radius" of K in the direction of θ and call it $r(\theta)$. This "directional radius" is a function on the sphere. Our aim is to find (say) a 10-dimensional slice of the sphere on which $r(\theta)$ is almost constant. In such a slice, the body K looks like a ball, since its radius hardly varies.

The fact that K is convex means that the function r cannot change too rapidly as we move around the sphere: if two directions are close together, then the radius of K must be about the same in these two directions. Now we apply the geometric deviation principle to conclude that the radius of K is roughly the same on almost the entire sphere: the radius is close to its average (or median) value for all but a small fraction of the possible directions. That means that we have plenty of room in which to go looking for a slice on which the radius is almost constant—we just have to choose a slice that avoids the small bad regions. It can be shown that this happens if we choose the slice at random from among all possible slices. The fact that most of the sphere consists of good regions means that a random slice has a good chance of falling into a good region.

Dvoretzky's theorem can be recast as a statement about the behavior of the entire body K , rather than just its sections, by using the Minkowski sums defined in the previous section. The statement is that if K is a convex body in n dimensions, then there is a family of m rotations K_1, K_2, \dots, K_m of K whose Minkowski sum $K_1 + \dots + K_m$ is approximately a ball, where the number m is significantly smaller than the dimension n . Recently, Milman and Schechtman realized that the smallest number m that would work could be described almost exactly, in terms of relatively simple properties of the body K , despite the apparently enormous complexity of the choice of rotations available.

For some n -dimensional convex sets, it is possible to create a ball with many fewer than n rotations. In the late 1970s Kašin discovered that if K is the cube, then just two rotations K_1 and K_2 are enough to produce something approximating a ball, even though the cube itself is extremely far from spherical. In two dimensions it is not hard to work out which rotations are best: if we choose K_1 to be a square and K_2 to be its rotation through 45° , then $K_1 + K_2$ is a regular octagon which is as close to a circle as we can get with just two squares. In higher dimensions it is extremely hard to describe which rotations to use. At present the only known method is to use randomly chosen rotations, even though the cube is as concrete and explicit an object as one ever meets in mathematics.

The strongest principle discovered to date showing that most bodies behave like balls is what is usually called the reverse Brunn-Minkowski inequality. This result was proved by Milman, building on ideas of his own and of Pisier and Bourgain. The Brunn-Minkowski inequality was stated earlier for sums of bodies. The reverse one has a number of different versions; the simplest is in terms of intersections. To begin with, if K is a body and B is a ball of the same volume, then the intersection of these two sets, the region that they have in common, is clearly of smaller volume. This obvious fact can be stated in a complicated way that looks like the Brunn-Minkowski inequality:

$$\text{vol}(K \cap B)^{1/n} \leq \text{vol}(K)^{1/n}. \quad (4)$$

If K is extremely long and thin, then whenever we intersect it with a ball of the same volume, we capture only a tiny part of K . So there is no possibility of reversing inequality (4) as it stands: no possibility of estimating the volume of $K \cap B$ from below. But if we are allowed to stretch the ball before intersecting it with K , the situation changes completely. A stretched

ball in n -dimensional space is called an ellipsoid (in two dimensions it is just an ellipse). The reverse Brunn–Minkowski inequality states that for every convex body K , there is an ellipsoid \mathcal{E} of the same volume for which

$$\text{vol}(K \cap \mathcal{E})^{1/n} \geq \alpha \text{vol}(K)^{1/n},$$

where α is a fixed positive number.

There is a widespread (but not quite universal) belief that an apparently much stronger principle is true: that if we are allowed to enlarge the ellipsoid by a factor of (say) 10, then we can ensure that it includes half the volume of K . In other words, for every convex body, there is an ellipsoid of roughly the same size which contains half of K . Such a statement flies in the face of our intuition about the huge variety of shapes in high dimensions, but there are some good reasons to believe it.

Since the Brunn–Minkowski inequality has a reverse form, it is natural to ask whether the isoperimetric inequality also does. The isoperimetric inequality guarantees that sets cannot have a surface that is too small. Is there a sense in which bodies cannot have too large a surface area? The answer is yes, and indeed a rather precise statement can be made. Just as in the case of the Brunn–Minkowski inequality, we have to take into account the possibility that our body could be long and thin and so have small volume but very large surface. So we have to start by applying a linear transformation that stretches the body in certain directions (but does not bend the shape). For example, if we start with a triangle, we first transform it into an *equilateral* triangle and then measure its surface and its volume. Once we have transformed our body as best we can, it turns out that we can specify precisely which convex body has the largest surface for a given volume. In two dimensions it is the triangle, in three it is the tetrahedron, and in n dimensions it is the natural analogue of these: the n -dimensional convex set (called a simplex) which has $n + 1$ corners. The fact that this set has the largest surface was proved by the present author using an inequality from harmonic analysis discovered by Brascamp and Lieb; the fact that the simplex is the only convex set with maximal surface (in the sense described) was proved by Barthe.

In addition to geometric deviation principles, two other methods played a central role in the modern development of high-dimensional geometry; these methods grew out of two branches of probability theory. One is the study of sums of random points in NORMED SPACES [III.64] and how big they are, which

provides important geometrical information about the spaces themselves. The other, the theory of Gaussian processes, depends upon a detailed understanding of how to cover sets in high-dimensional space efficiently with small balls. This issue may sound abstruse but it addresses a fundamental problem: how to measure (or estimate) the complexity of a geometric object. If we know that our object can be covered by one ball of radius 1, ten balls of radius $\frac{1}{2}$, fifty-seven balls of radius $\frac{1}{4}$, and so on, then we have a good idea of how complicated the object can be.

The modern view of high-dimensional space has revealed that it is at once much more complicated than was previously thought and at the same time in other ways much simpler. The first of these is well illustrated by the solution of a problem posed by Borsuk in the 1930s. A set is said to have diameter at most d if no two points in the set are further than d from each other. In connection with his work in topology, Borsuk asked whether every set of diameter 1 in n -dimensional space could be broken into $n + 1$ pieces of smaller diameter. In two and three dimensions this is always possible, and as late as the 1960s it was expected that the answer should be yes in all dimensions. However, a few years ago, Kahn and Kalai showed that in n dimensions it might require something like $e^{\sqrt{n}}$ pieces, enormously more than $n + 1$.

On the other hand, the simplicity of high-dimensional space is reflected in a fact discovered by Johnson and Lindenstrauss: if we pick a configuration of n points (in whatever dimension we like), we can find an almost perfect copy of the configuration sitting in a space of dimension much smaller than n : roughly the logarithm of n . In the last few years this fact has found applications in the design of computer algorithms, since many computational problems can be phrased geometrically and become much simpler if the dimension involved is small.

6 Deviation in Probability

If you toss a fair coin repeatedly, you expect that heads will occur on roughly half the tosses, and tails on roughly half. Moreover, as the number of tosses increases, you expect the proportion of heads to get closer and closer to $\frac{1}{2}$. The number $\frac{1}{2}$ is called the *expected number* of heads per toss. The number of heads yielded by a given toss is either 1 or 0, with equal probability, so the expected number of heads is the average of these, namely $\frac{1}{2}$.

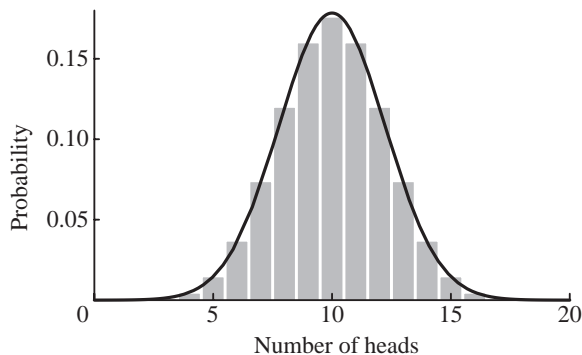


Figure 10 Twenty tosses of a fair coin.

The crucial unspoken assumption that we make about the tosses of the coin is that they are *independent*: that the outcomes of different tosses do not influence one another. (Independence and other basic probabilistic concepts are discussed in PROBABILITY DISTRIBUTIONS [III.73].) The coin-tossing principle, or its generalization to other random experiments, is called the *strong law of large numbers*. The average of a large number of independent repetitions of a random quantity will be close to the expected value of the quantity.

The strong law of large numbers for coin tosses is fairly simple to demonstrate. The general form, which applies to much more complicated random quantities, is considerably more difficult. It was first established by KOLMOGOROV [VI.88] in the early part of the twentieth century.

The fact that averages accumulate near the expected value is certainly useful to know, but for most purposes in statistics and probability theory it is vital to have more detailed information. If we focus our attention near the expected value, we may ask how the average is distributed around this number. For example, if the expected value is $\frac{1}{2}$, as for coin tossing, we might ask, what is the chance that the average is as large as 0.55 or as small as 0.42? We want to know how likely it is that our average number of heads will deviate from the expected value by a given amount.

The bar chart in figure 10 shows the probabilities of obtaining each of the possible numbers of heads, with twenty tosses of a coin. The height of each bar shows the chance that the corresponding number of heads will occur. As we would expect from the strong law of large numbers, the taller bars are concentrated near the middle. Superimposed upon the chart is a curve that plainly approximates the probabilities quite well.

This is the famous “bell-shaped” or “normal” curve. It is a shifted and rescaled copy of the so-called *standard normal curve*, whose equation is

$$y = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right). \quad (5)$$

The fact that the curve approximates coin-tossing probabilities is an example of the most important principle in probability theory: the *central limit theorem*. This states that whenever we add up a large number of small independent random quantities, the result has a distribution that is approximated by a normal curve.

The equation of the normal curve (5) can be used to show that if we toss a coin n times, then the chance that the proportion of heads deviates from $\frac{1}{2}$ by more than ε is at most $e^{-2n\varepsilon^2}$. This closely resembles the geometric deviation estimate (3) from section 4. This resemblance is not coincidental, although we are still far from a full understanding of when and how it applies.

The simplest way to see why a version of the central limit theorem might apply to geometry is to replace the toss of a coin by a different random experiment. Suppose that we repeatedly select a random number between -1 and 1 , and that the selection is *uniform* in the sense described in section 4. Let the first n selections be the numbers x_1, x_2, \dots, x_n . Instead of thinking of them as independent random choices, we can consider the point (x_1, \dots, x_n) as a randomly chosen point inside the cube that consists of all points whose coordinates lie between -1 and 1 . The expression $(1/\sqrt{n}) \sum_{i=1}^n x_i$ measures the distance of the random point from a certain $(n-1)$ -dimensional “plane,” which consists of all points whose coordinates add up to zero (the two-dimensional case is shown in figure 11). So the chance that $(1/\sqrt{n}) \sum_{i=1}^n x_i$ deviates from its expected value, 0, by more than ε is the same as the chance that a random point of the cube lies a distance of more than ε from the plane. This *chance* is proportional to the *volume* of the set of points that are more than ε from the plane: the set shown shaded in figure 11. When we discussed the geometric deviation principle, we estimated the volume of the set of points which were more than ε away from a set C which occupied half the ball. The present situation is really the same, because each part of the shaded set consists of those points that are more than ε away from whichever half of the cube lies on the other side of the plane.

Arguments akin to the central limit theorem show that if we cut the cube in half with a plane, then the set of points which lie more than a distance ε from one of

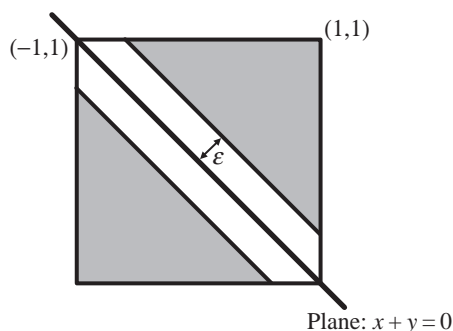


Figure 11 A random point of the cube.

the halves has volume no more than $e^{-\varepsilon^2}$. This statement is different from, and apparently much weaker than, the one we obtained for the ball (3) because the factor of n is missing from the exponent. The estimate implies that if you take any plane through the center of the cube, then most points in the cube will be at a distance of less than 2 from it. If the plane is parallel to one of the faces of the cube, this statement certainly is weak, because *all* of the cube is within distance 1 of the plane. The statement becomes significant when we consider planes like the one in figure 11. Some points of the cube are at a distance of \sqrt{n} from this “diagonal” plane, but still, the overwhelming majority of the cube is very much closer. Thus, the estimates for the cube and the ball contain essentially the same information; what is different is that the cube is bigger than the ball by a factor of about \sqrt{n} .

In the case of the ball we were able to prove a deviation estimate for *any* set occupying half the ball, not just the special sets that are cut off by planes. Towards the end of the 1980s Pisier found an elegant argument that showed that the general case works for the cube as well as for the ball. Among other things, the argument uses a principle which goes back to the early days of large-deviation theory in the work of Donsker and Varadhan.

The theory of large deviations in probability is now highly developed. In principle, more or less precise estimates are known for the probability that a sum of independent random variables deviates from its expectation by a given amount, in terms of the original distribution of the variables. In practice, the estimates involve quantities that may be difficult to compute, but there are sophisticated methods for doing this. The theory has numerous applications within probability and statistics, computer science, and statistical physics.

One of the most subtle and powerful discoveries of this theory is Talagrand's deviation inequality for product spaces, discovered in the mid 1990s. Talagrand himself has used this to solve several famous problems in combinatorial probability and to obtain striking estimates for certain mathematical models in particle physics. The full inequality of Talagrand is somewhat technical and is difficult to describe geometrically. However, the discovery had a precursor which fits perfectly into the geometric picture and which captures at least one of the most important ideas.² We look again at random points in the cube but this time the random point is not chosen uniformly from within the cube. As before, we choose the coordinates x_1, x_2, \dots, x_n of our random point *independently* of one another, but we do not insist that each coordinate is chosen *uniformly* from the range between -1 and 1 . For example, it might be that x_1 can take only the values $1, 0$, or -1 , each with probability $\frac{1}{3}$, that x_2 can take only the values 1 or -1 each with probability $\frac{1}{2}$, and perhaps that x_3 is chosen uniformly from the entire range between -1 and 1 . What matters is that the choice of each coordinate has no effect on the choice of any others.

Any sequence of rules that dictates how we choose each coordinate determines a way of choosing a random point in the cube. This in turn gives us a way of measuring a kind of volume for subsets of the cube: the “volume” of a set A is the chance that our random point is selected from A . This way to measure volume might be very different from the usual one; among other things, an individual point might have nonzero volume.

Now suppose that C is a *convex* subset of the cube and that its “volume” is $\frac{1}{2}$, in the sense that our random point will be selected from C with probability $\frac{1}{2}$. Talagrand's inequality says that the chance that our random point will lie a distance of more than ε from C is less than $2e^{-\varepsilon^2/16}$. This statement looks like the deviation estimate for the cube except that it refers only to convex sets C . But the crucial *new* information that makes the estimate and its later versions important is that we are allowed to choose our random point in so many different ways.

This section has described deviation estimates in probability theory that have a geometric flavor. For the cube, we are able to show that if C is any set occupying half the cube, then almost the entire cube is close to C . It would be extremely useful to know the same thing for

2. This precursor evolved from an original argument of Talagrand via an important contribution of Johnson and Schechtman.

convex sets more general than the cube. There are some other highly symmetric sets for which we do know it, but the most general possible statement of this type seems to be beyond our current methods. One potential application, which comes from theoretical computer science, is to the analysis of random algorithms for volume calculation. The problem may sound specialized, but it arises in LINEAR PROGRAMMING [III.86] (which alone is sufficient reason to justify the expenditure of enormous effort) and in the numerical estimation of integrals. In principle, one can calculate the volume of a set by laying over it a very fine grid, and counting how many grid points fall into the set. In practice, if the dimension is large, the number of grid points will be so astronomically huge that no computer has a chance of performing the count.

The problem of calculating the volume of a set is essentially the same as the problem of choosing a point at random within the set, roughly as we saw in section 4. So the aim is to select a random point without identifying a huge number of possible points to select from. At present, the most effective way of generating a random point in a convex set is to carry out a random walk within the set. We perform a sequence of small steps whose directions are chosen randomly and then select the point that we have reached after a fairly large number of steps, in the hope that this point has roughly the correct chance of falling into each part of the set. For the method to be effective, it is essential that the random walk quickly visits points all over the set: that it does not get stuck for a long time in, say, half of the set. In order to guarantee this *rapid mixing*, as it is called, we need an isoperimetric principle or deviation principle. We need to know that each half of our set has a large boundary, so that there is a good chance that our random walk will cross the boundary quickly and land in the other half of our set.

In a series of papers published over the last ten years, Applegate, Bubley, Dyer, Frieze, Jerrum, Kannan, Lovasz, Montenegro, Simonovits, Vempala, and others have found very efficient random walks for sampling from a convex set. A geometric deviation principle of the kind alluded to above would make it possible to estimate the efficiency of these random walks almost perfectly.

7 Conclusion

The study of high-dimensional systems has become increasingly important in the last few decades. Prac-

tical problems in computing frequently lead to high-dimensional questions, many of which can be posed geometrically, while many models in particle physics are automatically high-dimensional because it is necessary to consider a huge number of particles in order to mimic large-scale phenomena in the real world. The literature in both these fields is vast but some general remarks can be made. The intuition that we gain from low-dimensional geometry leads us wildly astray if we try to apply it in many dimensions. It has become clear that naturally occurring high-dimensional systems exhibit characteristics that we expect to arise in probability theory, even if the original system does not have an explicitly random element. In many cases these random characteristics are manifested as an isoperimetric or deviation principle, that is, a statement to the effect that large sets have large boundaries. In the classical theory of probability, independence assumptions can often be used to demonstrate deviation principles quite simply. For the very much more complicated systems that are studied today it is usually useful to have a geometric picture to accompany the probabilistic one. That way one can understand probabilistic deviation principles as analogues of the isoperimetric principle discovered by the ancient Greeks. This article has described the relationship between geometry and probability in just a few special cases. A very much more detailed picture is almost certainly waiting to be found. At present it seems to be just out of reach.

Further Reading

- Ball, K. M. 1997. An elementary introduction to modern convex geometry. In *Flavors of Geometry*, edited by Silvio Levy. Cambridge: Cambridge University Press.
- Bollobás, B. 1997. Volume estimates and rapid mixing. In *Flavors of Geometry*, edited by Silvio Levy. Cambridge: Cambridge University Press.
- Chavel, I. 2001. *Isoperimetric Inequalities*. Cambridge: Cambridge University Press.
- Dembo, A., and O. Zeitouni. 1998. *Large Deviations Techniques and Applications*. New York: Springer.
- Ledoux, M. 2001. *The Concentration of Measure Phenomenon*. Providence, RI: American Mathematical Society.
- Osserman, R. 1978. The isoperimetric inequality. *Bulletin of the American Mathematical Society* 84:1182-238.
- Pisier, G. 1989. *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge: Cambridge University Press.
- Schneider, R. 1993. *Convex Bodies: The Brunn-Minkowski Theory*. Cambridge: Cambridge University Press.

PUP: further reading added after proofreader received proof. Please check.

Part V

Theorems and Problems

V.1 The ABC Conjecture

The ABC conjecture, proposed by Masser and Oesterlé in 1985, is a bold and very general conjecture in number theory with a wide range of important consequences. The rough idea of the conjecture is that it is impossible for one number to be the sum of two others if all three numbers have many repeated prime factors and no two have a prime factor in common (which would then have to be shared by the third).

More precisely, one defines the *radical* of a positive integer n to be the product of all primes that divide n , with each distinct prime included just once. For instance, $3960 = 2^3 \times 3^2 \times 5 \times 11$, so its radical is $2 \times 3 \times 5 \times 11 = 330$. Let us write $\text{rad}(n)$ for the radical of n . The ABC conjecture asserts that for every positive real number ϵ there is a constant K_ϵ such that if a , b , and c are coprime integers and $a + b = c$, then $c < K_\epsilon \text{rad}(abc)^{1+\epsilon}$.

To get a feel for the meaning of this conjecture, consider the Fermat equation $x^r + y^r = z^r$ (see FERMAT'S LAST THEOREM [V.12]). If three positive integers x , y , and z solve the equation, then we can divide through by any common factors they might have and obtain a solution for which x , y , and z , and hence their r th powers, are coprime. Set $a = x^r$, $b = y^r$, and $c = z^r$. Then

$$\text{rad}(abc) = \text{rad}(xyz) \leq xyz = (abc)^{1/r} \leq c^{3/r},$$

where the last inequality follows from the fact that c is greater than either a or b . If we set ϵ to be $\frac{1}{6}$, then the ABC conjecture gives us a constant K such that c cannot be more than $K(c^{3/r})^{7/6} = Kc^{7/2r}$. If $r \geq 4$ then the power $7/2r$ is less than 1, so the Fermat equation can have at most finitely many solutions with x , y , and z coprime.

It is clear that this is just one of a huge number of consequences of a similar kind. For instance, we could deduce that there are only finitely many solutions of the equation $2^r + 3^s = x^2$, since the radical

of $2^r 3^s x^2$ is $6x$, which is considerably smaller than x^2 . But the ABC conjecture has other consequences that are less obvious, and more important, than this one. For instance, Bombieri has shown that the ABC conjecture implies ROTH'S THEOREM [V.25], Elkies has shown that it implies the MORDELL CONJECTURE [V.32], and Granville and Stark have shown that a strengthening of the ABC conjecture implies the nonexistence of Siegel zeros (these are defined in ANALYTIC NUMBER THEORY [IV.2]). It is also equivalent to strong forms, as yet unproven, of a famous theorem of Baker in transcendence theory, and of the theorem of Wiles about MODULAR FORMS [III.61] that implies Fermat's last theorem.

The ABC conjecture is discussed further in COMPUTATIONAL NUMBER THEORY [IV.3].

V.2 The Atiyah-Singer Index Theorem

Nigel Higson and John Roe

1 Elliptic Equations

The Atiyah-Singer index theorem is concerned with the existence and uniqueness of solutions to linear partial differential equations of *elliptic type*. To understand this concept, consider the two equations

$$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} = 0 \quad \text{and} \quad \frac{\partial f}{\partial x} + i \frac{\partial f}{\partial y} = 0.$$

They differ only by the factor $i = \sqrt{-1}$, but their solutions nevertheless have very different properties. Any function of the form $f(x, y) = g(x - y)$ is a solution to the first equation, but in the analogous general solution $g(x + iy)$ of the second equation, g must be a HOLOMORPHIC FUNCTION [I.3 §5.6] of the *complex* variable $z = x + iy$, and it was already known in the nineteenth century that such functions are very special. For example, the first equation has an infinite-dimensional set of bounded solutions, but LIOUVILLE'S

THEOREM [I.3 §5.6] in complex analysis asserts that the only bounded solutions of the second equation are the constant functions.

The differences between the solutions of the two equations can be traced to the differences between the *symbols* of the equations, which are the polynomials in real variables ξ, η that are obtained by substituting $i\xi$ for $\partial/\partial x$ and $i\eta$ for $\partial/\partial y$. Thus the symbols of the two equations above are

$$i\xi + i\eta \quad \text{and} \quad i\xi - \eta,$$

respectively. An equation is said to be *elliptic* if its symbol is zero only when $\xi = \eta = 0$; thus, the second equation is elliptic but the first is not. The fundamental *regularity theorem*, which is proved using FOURIER ANALYSIS [III.27], states that an elliptic partial differential equation (subject to suitable boundary conditions, if needed) has a finite-dimensional solution space.

2 Topology of Elliptic Equations and the Fredholm Index

Consider now the general first-order linear partial differential equation

$$a_1 \frac{\partial f}{\partial x_1} + \cdots + a_n \frac{\partial f}{\partial x_n} + b f = 0,$$

in which f is a vector-valued function and the coefficients a_j and b are complex matrix-valued functions. It is *elliptic* if its *symbol*

$$i\xi_1 a_1(x) + \cdots + i\xi_n a_n(x)$$

is an invertible matrix for every nonzero vector $\xi = (\xi_1, \dots, \xi_n)$ and every x . The regularity theorem applies in this generality, and it allows us to form the *Fredholm index* of an elliptic equation (with suitable boundary conditions), which is the number of linearly independent solutions of the equation minus the number of linearly independent solutions of the *adjoint equation*

$$-\frac{\partial}{\partial x_1}(a_1^* f) - \cdots - \frac{\partial}{\partial x_n}(a_n^* f) + b^* f = 0.$$

The reason for introducing the Fredholm index is that it is a *topological invariant* of elliptic equations. This means that continuous variations in the coefficients of an elliptic equation leave the Fredholm index unchanged. (By contrast, the number of linearly independent solutions of an equation can vary as the coefficients of the equation vary.) The Fredholm index is therefore constant on each connected component of the set of all elliptic equations, and this raises the prospect of using topology to determine the structure of the set of all elliptic equations as an aid to

computing the Fredholm index. This observation was made by Gelfand in the 1950s. It lies at the root of the Atiyah-Singer index theorem.

3 An Example

To see in more detail how topology can be used to determine the Fredholm index of an elliptic equation, let us look at a specific example. Consider elliptic equations for which the coefficients $a_j(x)$ and $b(x)$ are *polynomial* functions of x , with a_j of degree $m-1$ or less and b of degree m or less. The expression

$$i\xi_1 a_1(x) + \cdots + i\xi_n a_n(x) + b(x)$$

is then a polynomial in both x and ξ of degree m or less. Let us strengthen the hypothesis of ellipticity by assuming that the terms in this expression that have degree exactly m (jointly in x and ξ) define an invertible matrix whenever *either* x *or* ξ is nonzero. Let us also agree to consider only solutions f of the equation or its adjoint that are *square-integrable*, which means that

$$\int |f(x)|^2 dx < \infty.$$

All these extra hypotheses are types of boundary conditions (the behaviors of the equation and its solutions at infinity are controlled), and collectively they imply that the Fredholm index is well-defined.

A simple example is the equation

$$\frac{df}{dx} + x f = 0. \quad (1)$$

The general solution to this ordinary differential equation is the one-dimensional space of multiples of the square-integrable function $e^{-x^2/2}$. By contrast, the solutions of the adjoint equation

$$-\frac{df}{dx} + x f = 0$$

are multiples of the function $e^{+x^2/2}$, which is not square-integrable. Thus the index of this differential equation is equal to 1.

Returning to the general equation, the terms of degree m in

$$i\xi_1 a_1(x) + \cdots + i\xi_n a_n(x) + b(x)$$

determine a map from the unit sphere in (x, ξ) -space to the set $GL_k(\mathbb{C})$ of invertible $k \times k$ complex matrices. Moreover, every such map comes from an elliptic equation (possibly of a more general type than we have discussed up to now, but an equation to which the basic regularity theorem guaranteeing the existence of the

Fredholm index applies). It therefore becomes important to determine the topological structure of the space of all maps from the sphere S^{2n-1} into $GL_k(\mathbb{C})$.

A remarkable theorem of Bott provides the answer. The *Bott periodicity theorem* associates an integer, which we shall call the *Bott invariant*, with each map $S^{2n-1} \rightarrow GL_k(\mathbb{C})$. Furthermore, Bott's theorem asserts that, provided that $k \geq n$, one such map can be continuously deformed into another if and only if the Bott invariants of the two maps agree. In the special case $n = k = 1$, where we are dealing with maps from the one-dimensional circle into the nonzero complex numbers, or in other words closed paths in \mathbb{C} that do not pass through the origin, the Bott invariant is just the classical *winding number*, which measures the number of times such a path winds around the origin. We may therefore regard the Bott invariant as a generalized winding number.

The index theorem for equations of the type that we are considering in this section asserts that the Fredholm index of an elliptic equation is equal to the Bott invariant of its symbol. For instance, in the case of the simple example (1) considered above, the symbol $i\xi + x$ corresponds to the identity map from the unit circle in (x, ξ) -space to the unit circle in \mathbb{C} . Its winding number is equal to 1, in agreement with our computation of the index.

The proof of the index theorem depends strongly on Bott periodicity and proceeds as follows. Because elliptic equations are classified topologically by the Bott invariant, and because the Bott invariant and the Fredholm index have analogous algebraic properties, one need only verify the theorem in a single example: that corresponding to a symbol with Bott invariant 1. It turns out that this *Bott generator* can be represented by an n -dimensional generalization of our example (1), and a computation in this case completes the proof.

4 Elliptic Equations on Manifolds

It is possible to define elliptic equations not just for functions f of n variables, but also for functions defined on a MANIFOLD [I.3 §6.9]. Particularly accessible to analysis are the elliptic equations on *closed* manifolds, that is, on manifolds that are finite in extent and that have no boundary. For closed manifolds it is not necessary to specify any boundary conditions in order to obtain the basic regularity theorem for elliptic equations (after all, there is no boundary). As a result, every elliptic partial differential equation on a closed manifold has a Fredholm index.

The Atiyah-Singer index theorem concerns elliptic equations on closed manifolds and it has roughly the same form as the index theorem that we studied in the previous section. One builds out of the symbol an invariant called the *topological index*, which generalizes the Bott invariant. The Atiyah-Singer index theorem then asserts that the topological index of an elliptic equation is equal to the Fredholm or *analytical* index of the equation. The proof has two stages. In the first, theorems are proved that allow one to transform an elliptic equation on a general manifold into an elliptic equation on a sphere without changing the topological or analytical indices. For example, it may be shown that two elliptic equations on different manifolds that are the common "boundary" of an elliptic equation on a manifold of one higher dimension must have the same topological and analytical indices. In the second stage of the proof the Bott periodicity theorem and an explicit computation are applied to identify the topological and analytical indices of elliptic equations on spheres. Throughout both stages, an important tool is *K-THEORY* [IV.6 §6], which is a branch of algebraic topology invented by Atiyah and Hirzebruch.

Although the proof of the Atiyah-Singer index theorem makes use of *K-theory*, the final result can be translated into terms that do not mention *K-theory* explicitly. In this way one obtains an index formula roughly like this:

$$\text{index} = \int_M I_M \cdot \text{ch}(\sigma).$$

The term I_M is a DIFFERENTIAL FORM [III.16] determined by the CURVATURE [III.80] of the manifold M on which the equation is defined. The term $\text{ch}(\sigma)$ is a differential form obtained from the symbol of the equation.

5 Applications

In order to prove the index theorem, Atiyah and Singer were obliged to study a very broad class of generalized elliptic equations. However, the applications they first had in mind were related to the simple equation with which we began this article. Solutions of the equation

$$\frac{\partial f}{\partial x} + i \frac{\partial f}{\partial y} = 0$$

are precisely the analytic functions of the complex variable $z = x + iy$. There is a counterpart to this equation on any RIEMANN SURFACE [III.81], and the Atiyah-Singer index formula, applied in this instance, is equivalent to a foundational result about the geometry of surfaces called the RIEMANN-ROCH THEOREM [V.34]. The Atiyah-Singer index theorem then gives a means to generalize

PUP: Tim thinks that the small effort involved by the reader in figuring out what to do here is well worth it and that it should stay as it is. OK?

the Riemann–Roch theorem to a COMPLEX MANIFOLD [III.6 §2] of any dimension.

The Atiyah–Singer index theorem also has important applications outside of complex geometry. The simplest example involves the elliptic equation $d\omega + d^*\omega = 0$, concerning differential forms on a manifold M . The Fredholm index may be identified with the *Euler characteristic* of M , which is the alternating sum of the numbers of r -dimensional cells in a cell decomposition of M . For two-dimensional manifolds, the Euler characteristic is the familiar quantity $V - E + F$. In the two-dimensional case, the index theorem reproduces the Gauss–Bonnet theorem, which asserts that the Euler characteristic is a multiple of the total Gaussian curvature.

Even in this simple case, the index theorem can be used to produce topological restrictions on the ways a manifold can curve. Many important applications of the index theorem proceed in the same direction. For example, Hitchin used a more refined application of the Atiyah–Singer index theorem to show that there is a nine-dimensional manifold that is homeomorphic to the sphere despite not being positively curved in even the weakest sense. (By contrast, the usual sphere is positively curved in the strongest possible sense.)

Further Reading

- Atiyah, M. F. 1967. Algebraic topology and elliptic operators. *Communications in Pure and Applied Mathematics* 20:237–49.
- Atiyah, M. F., and I. M. Singer. 1968. The index of elliptic operators. I. *Annals of Mathematics* 87:484–530.
- Hirzebruch, F. 1966. *Topological Methods in Algebraic Geometry*. New York: Springer.
- Hitchin, N. 1974. Harmonic spinors. *Advances in Mathematics* 14:1–55.

V.3 The Banach–Tarski Paradox

T. W. Körner

The Banach–Tarski paradox states that we can decompose the three-dimensional unit sphere into a finite number of disjoint pieces which we can then translate and rotate in such a way that they remain disjoint but now their union consists of two copies of our original sphere.

Such a result seems impossible at first sight, and indeed it contradicts the naive assumption that one can consistently assign a finite volume to every bounded

set. In other words, it shows that one cannot assign volumes to *all* bounded sets in such a way that these volumes are unaffected by translation and rotation, that the volume of a union of two disjoint sets is the sum of the volumes of the two sets, and that the volume of the unit sphere is greater than zero. However, if we drop this naive assumption, then the paradox disappears. Since there is no genuine paradox, we shall refer to the Banach–Tarski *construction*.

The Banach–Tarski construction is a descendant of an older construction due to Vitali, which concerns area rather than volume. Let us write l_θ for the line segment in \mathbb{R}^2 that is given in polar coordinates by

$$l_\theta = \{(r, \theta) : 0 < r \leq 1\}.$$

Note that the union of all such segments is the punctured unit disk D_* (that is to say, the unit disk with the origin removed). We say that l_θ and l_ϕ belong to the same equivalence class if $\theta - \phi$ is a rational multiple of π , and we consider a set E that is the union of a set of l_θ containing *exactly one representative* from each equivalence class.

The rationals are COUNTABLE [III.11], so we can enumerate the rationals x with $0 \leq x < 1$ as a sequence x_1, x_2, \dots . If we write

$$E_n = \{l_{\theta+2\pi x_n} : l_\theta \in E\},$$

then each E_n is obtained from E by a rotation about the origin (through an angle $2\pi x_n$), the E_n are disjoint (as E contains only one representative from each equivalence class), and the union of the E_n is D_* (as E contains a representative from each equivalence class).

Now take D_* and split it into the set F consisting of the union of the sets E_{2n} and the set G consisting of the union of the sets E_{2n+1} . Each E_{2n} can be rotated to E_n , and the union of the E_n gives us D_* . Similarly, each E_{2n+1} can be rotated to E_n , and the union of the E_n gives us D_* again. Thus the punctured unit disk can be split into a countable set of disjoint pieces (all obtained by rotation of one particular set) which can be rotated and translated to form disjoint sets whose union is two copies of D_* .

Vitali’s construction makes use of THE AXIOM OF CHOICE [III.1] (because we chose one representative from each equivalence class), and the same is true for the Banach–Tarski construction. Solovay showed that if we reject the axiom of choice, then there are MODELS OF SET THEORY [IV.22 §3] in which it is possible to assign a volume to all bounded sets in \mathbb{R}^3 in a consistent way. However, most mathematicians would agree

that the natural moral to draw from our discussion is that when we define volume we should consider only a restricted collection of sets.

The Banach-Tarski construction is also closely related to our final example, which requires a little group theory. To introduce this example of bad behavior, we first consider an example of good behavior. Suppose that $f : \mathbb{R} \rightarrow \mathbb{R}$ is a reasonable function with $f(x) \geq 0$ and $f(x+1) = f(x)$ for all x (thus, f is positive and periodic with period 1). Suppose that there existed real numbers s, t, u, v such that

$$f(x+s) + f(x+t) - f(x+u) - f(x+v) \leq -1 \quad (1)$$

for all x . Since $\int_0^1 f(x+w) dx = \int_0^1 f(x) dx$ for all w , integrating both sides of (1) from 0 to 1 would give

$$0 \leq \int_0^1 (-1) dx = -1,$$

which is impossible. Thus (1) cannot hold.

Now consider the FREE GROUP [IV.10 §2] G generated by a and b (that is to say, the group generated by a and b where no nontrivial relations hold between a and b). Every element of G can be written in shortest form as the product of a sequence, each term of which is a, a^{-1}, b , or b^{-1} . Define $F(x) = 1$ if $x = e$ or the shortest form of x ends with a or a^{-1} , and set $F(x) = 0$ otherwise. We see that $F(x) \geq 0$ for all $x \in G$, and the reader can check, by going through cases, that

$$F(xb) + F(xab) - F(xa^{-1}) - F(xb^{-1}a) \leq -1 \quad (2)$$

for all $x \in G$. The averaging argument that enabled us to show that (1) was false for \mathbb{R} must fail for G since (2) is, in fact, true. If there is no averaging argument, then there can be no appropriate universal integral and no appropriate universal “volume” in G .

This example bears a clear family resemblance to the “paradoxes” discussed earlier. If we consider the group $SO(3)$ of rotations in three dimensions, then (unless specific conditions hold) there is no nontrivial group relation between two generally chosen rotations A and B about two generally chosen axes. Thus $SO(3)$ contains a copy of the group G considered in the previous paragraph. The Banach-Tarski construction is a modification of a construction of Hausdorff that exploits this fact.

There is a beautiful account of all these matters in *The Banach-Tarski Paradox* by Stan Wagon (Cambridge University Press, Cambridge, UK, 1993).

V.4 The Birch-Swinnerton-Dyer Conjecture

Given an ELLIPTIC CURVE [III.21], there is a natural way of defining a binary operation on its points, and this turns the elliptic curve into an ABELIAN GROUP [I.3 §2.1]. Moreover, the points on the curve with rational coordinates form a subgroup of this group. Mordell’s theorem tells us that this subgroup is finitely generated. (These results are described in RATIONAL POINTS ON CURVES AND THE MORDELL CONJECTURE [V.32].)

Every finitely generated Abelian group is isomorphic to a group of the form $\mathbb{Z}^r \times C_{n_1} \times C_{n_2} \times \cdots \times C_{n_k}$, where C_n stands for the cyclic group with n elements. The number r , which measures the maximum number of independent elements of this group that have infinite order, is called the *rank* of the elliptic curve. Mordell’s theorem implies that the rank of every elliptic curve is finite, but it does not tell us how to calculate it. That turns out to be an extraordinarily hard problem: in fact, so hard that it is considered a remarkable achievement of Birch and Swinnerton-Dyer even to have come up with a plausible conjecture about it.

Their conjecture relates the rank of an elliptic curve to a very different object associated with that curve: an L -FUNCTION [III.49]. This is a function with properties similar to those of the RIEMANN ZETA FUNCTION [IV.2 §3], but it is defined in terms of a series of numbers $N_2(E), N_3(E), N_5(E), \dots$, one for each prime p ; the number $N_p(E)$ is the number of points on the elliptic curve when it is considered as a curve over the FIELD [I.3 §2.2] with p elements. One of the properties of the L -function of E is that it is HOLOMORPHIC [I.3 §5.6]. (The fact that it can be extended to a holomorphic function everywhere on the complex plane is very far from obvious: it follows from the fact that all elliptic curves are modular. See FERMAT’S LAST THEOREM [V.12].) Birch and Swinnerton-Dyer conjectured that the rank of the group associated with the elliptic curve is equal to the order of the zero of its L -function at 1. (If the L -function does not take the value 0 at 1, then this order is defined to be 0.) This can be thought of as a sophisticated LOCAL-TO-GLOBAL PRINCIPLE [III.53], in that it relates the rational solutions to the equation for the elliptic curve to the solutions mod p for each prime p .

Another remarkable feature of the conjecture is that far less was known about elliptic curves when Birch and Swinnerton-Dyer made it. Now there are many reasons to find it plausible, but then it was much more of a

leap in the dark: they based it on numerical evidence gleaned from computations of $N_p(E)$ for several elliptic curves and many primes p . In other words, they did not calculate the orders of zeros of L -functions of various elliptic curves, since that was too hard, but guessed them based on approximations.

The Birch-Swinnerton-Dyer conjecture has been proved for curves with L -functions that have a zero of order 0 or 1 at 1, but a proof of the general case still appears to be a long way off. It is one of the problems for which the Clay Mathematics Institute offers a prize of a million dollars. For a further discussion of the problem and much more about its mathematical context, see ARITHMETIC GEOMETRY [IV.5].

V.5 Carleson's Theorem

Charles Fefferman

Carleson's theorem asserts that the FOURIER SERIES [III.27] of a function f in $L^2[0, 2\pi]$ converges almost everywhere. To understand this statement and appreciate its significance, let us follow the history of the subject, starting in the early nineteenth century. FOURIER's [VI.25] great idea was that "any" (complex-valued) function f on an interval such as $[0, 2\pi]$ can be expanded in what we would now call a *Fourier series*,

$$f(\theta) = \sum_{n=-\infty}^{\infty} a_n e^{in\theta}, \quad (1)$$

for suitable *Fourier coefficients* a_n . Fourier obtained the formula for the coefficients a_n , and proved that (1) holds in interesting special cases.

The next major advance was due to DIRICHLET [VI.36], who gave a formula for the N th partial sum $S_N f(\theta)$, which is defined to be

$$S_N f(\theta) = \sum_{n=-N}^N a_n e^{in\theta}. \quad (2)$$

Dirichlet realized that the precise meaning of (1) is that

$$\lim_{N \rightarrow \infty} S_N f(\theta) = f(\theta). \quad (3)$$

Dirichlet used his formula for $S_N f$ to prove that under certain circumstances (3) does indeed hold. For example, if f is a continuous increasing function on $[0, 2\pi]$, then it holds for every $\theta \in (0, 2\pi)$.

Decades later, DE LA VALLÉE POUSSIN [VI.67] discovered an example of a continuous function whose Fourier series diverges at a single point. More generally, given any countable set $E \subset [0, 2\pi]$, there exists a continuous function f whose Fourier series diverges at

every point of E , a result that appears to restrict quite considerably the circumstances under which Fourier's original vision is valid.

The work of LEBESGUE [VI.72] led to fundamental progress in Fourier analysis and a significant change of viewpoint. We first sketch Lebesgue's ideas and then trace their impact on Fourier analysis.

Lebesgue sought to define a notion of integration that could be applied to all but the most pathological non-negative functions F on $[0, 2\pi]$. He began by defining the MEASURE [III.57] of a set $E \subset [0, 2\pi]$. Loosely speaking, the measure of E , written $\mu(E)$, is "what the set E would weigh" if the interval $[0, 2\pi]$ were made of wire weighing one gram per centimeter. For instance, the measure of an interval (a, b) is equal to its length $b - a$. Certain sets E have measure zero, e.g., countable sets, or the CANTOR SET [III.17]; sets of measure zero are regarded as negligibly small.

Using his notion of measure, Lebesgue defined the *Lebesgue integral* $\int_0^{2\pi} F(\theta) d\theta$ for the "measurable" functions $F \geq 0$ on $[0, 2\pi]$. All but the most pathological functions are measurable, but $\int_0^{2\pi} F(\theta) d\theta$ may be infinite if F is too big. For example, if $F(\theta) = 1/\theta$ for $\theta \in (0, 2\pi]$, then the integral of F is infinite.

Finally, for any number $p \geq 1$, the *Lebesgue space* $L^p[0, 2\pi]$ consists of all measurable functions f on $[0, 2\pi]$ that are not too big, in the sense that $\int_0^{2\pi} |f(\theta)|^p d\theta$ is finite. (See FUNCTION SPACES [III.29] for a slight, technical correction to this definition.)

We now turn to the impact of Lebesgue's theory on Fourier analysis. The Lebesgue space $L^2[0, 2\pi]$, which is also a HILBERT SPACE [III.37], plays a fundamental role. If f belongs to $L^2[0, 2\pi]$, then its Fourier coefficients a_n are such that

$$\sum_{n=-\infty}^{\infty} |a_n|^2 < \infty. \quad (4)$$

Conversely, any sequence of complex numbers a_n ($-\infty < n < \infty$) satisfying (4) arises as the sequence of Fourier coefficients of a function f in $L^2[0, 2\pi]$. Moreover, the size of a function f and its Fourier coefficients a_n are related by the *Plancherel formula*:

$$\frac{1}{2\pi} \int_0^{2\pi} |f(\theta)|^2 d\theta = \sum_{n=-\infty}^{\infty} |a_n|^2.$$

Finally, the partial sums $S_N f$ (see (2)) converge to the function f in the L^2 -norm. In other words,

$$\int_0^{2\pi} |S_N f(\theta) - f(\theta)|^2 d\theta \rightarrow 0 \quad (5)$$

as N tends to infinity. This gives us a precise sense in which the function f is the sum of its Fourier series.

PUP: Tim thinks the 'basic question' referred to at the end of the article is implicit in a couple of places (at the end of the paragraph that starts 'Decades later' and at the end of the later paragraph that starts 'However, it would still be nice') and that it is all clear enough as it is. OK?

Thus, we have justified Fourier's formula (1) by reinterpreting it as the statement (5) rather than using the more obvious interpretation of (3).

However, it would still be nice to know to what extent the original, more straightforward interpretation can be justified. In 1906, Luzin conjectured that if f is any function in $L^2[0, \pi]$, then

$$\lim_{N \rightarrow \infty} S_N f(\theta) = f(\theta) \quad (6)$$

for all θ outside a set of measure zero. When this holds, one says that the Fourier series of f converges *almost everywhere*. If Luzin's conjecture were true, it would validate Fourier's vision from the early nineteenth century.

For several decades it looked as if Luzin's conjecture might well be false. KOLMOGOROV [VI.88] constructed a function f in $L^1[0, 2\pi]$ whose Fourier series converges nowhere. Also, a theorem of Kolmogorov, Seliverstov, and Plessner, which asserted that $\lim_{N \rightarrow \infty} (S_N f(\theta) / \sqrt{\log N}) = 0$ almost everywhere when f is in $L^2[0, 2\pi]$, withstood all attempts at improvement for over thirty years.

It therefore came as a big surprise when Lennart Carleson proved in 1966 that Luzin's conjecture is true. The main point of Carleson's proof is to control the *Carleson maximal function*

$$C(f)(\theta) = \sup_{N \geq 1} |S_N f(\theta)|$$

by proving that

$$\mu(\{\theta \in [0, 2\pi] : C(f)(\theta) > \alpha\}) \leq \frac{A}{\alpha^2} \int_0^{2\pi} |f(\theta)|^2 d\theta \quad (7)$$

for all f in $L^2[0, 2\pi]$ and all $\alpha > 0$, where A is a constant independent of f and α . It is not hard to show that (7) implies Luzin's conjecture, but it is very hard to prove (7).

Shortly after Carleson's work, Hunt proved the almost-everywhere convergence of Fourier series of functions in $L^p[0, 2\pi]$ for any $p > 1$. Kolmogorov's counterexample shows that the result fails for $p = 1$.

Fourier analysis has been immensely useful in mathematics and its applications. (For a fuller discussion of this, see THE FOURIER TRANSFORM [III.27] and HARMONIC ANALYSIS [IV.11].) The theorems of Carleson and Hunt provide the sharpest known answer to the basic question that started the subject.

Acknowledgments. This work was partially supported by NSF grant #DMS-0245242.

V.6 Cauchy's Theorem

Cauchy's theorem asserts that if f is a HOLOMORPHIC FUNCTION [I.3 §5.6] defined on a simply connected domain (that is, an open set in \mathbb{C} with no holes), then the path integral of f around any closed curve that lies in the domain is zero. This theorem stands at the beginning of a remarkable series of results about holomorphic functions, such as the *residue theorem*, which is an extremely powerful way of calculating path integrals.

It follows fairly easily from Cauchy's theorem as stated above that the path integral of any holomorphic function around a closed curve C is a *topological invariant* of that curve, even when the domain in question is not simply connected. That is, the integral does not change if the curve is continuously deformed within the domain. This is just one indication of the geometric significance of the theorem, which is a vital tool for studying manifolds with complex structure, such as RIEMANN SURFACES [III.81] and KÄHLER MANIFOLDS [III.90 §3].

V.7 The Central Limit Theorem

The central limit theorem is a fundamental result in probability concerning sums of independent random variables. Let X_1, X_2, \dots be independent and suppose that they are identically distributed. Suppose also that they have mean 0 and variance 1. Then $X_1 + \dots + X_n$ has mean 0 and variance n . (The variance is n because the X_i are independent.) Therefore, $Y_n = (X_1 + \dots + X_n) / \sqrt{n}$ has mean 0 and variance 1. The central limit theorem states that, regardless of the distribution of the X_i , the random variable Y_n converges to a standard normal distribution. It is easy to deduce from this a similar result for random variables with any finite mean and variance. Details may be found in PROBABILITY DISTRIBUTIONS [III.73 §5].

V.8 The Classification of Finite Simple Groups

Martin W. Liebeck

A finite group G is said to be *simple* if its only normal subgroups are the identity subgroup and G itself. To some extent, simple groups play an analogous role in finite group theory to that of prime numbers in number theory: just as the only factors of a prime p are

1 and p itself, so the only factor groups of a simple group G are the identity group 1 and G itself. The analogy runs a bit deeper: just as every positive integer (greater than 1) is a product of a collection of primes, so every finite group is “built” from a collection of simple groups, in the following sense. Let H be a finite group, and choose a maximal normal subgroup H_1 of H (this means that H_1 is not the whole of H , and it is not contained in any larger normal subgroup that is not the whole of H); then choose a maximal normal subgroup H_2 of H_1 ; and so on. This gives a sequence of subgroups $1 = H_r < H_{r-1} < \cdots < H_1 < H_0 = H$, each one a maximal normal subgroup of the next, and, because of the maximality, each factor group $G_i = H_i/H_{i+1}$ is a simple group. It is in this sense that one says that H is built from the collection G_0, G_1, \dots, G_{r-1} of simple groups (although unlike the situation with prime numbers, there will in general be several different finite groups that are built from the same collection of simple groups).

At any rate, it is abundantly clear that simple groups lie at the heart of the theory of finite groups, and one of the driving forces of twentieth-century finite group theory was to study, and ultimately to classify completely, the finite simple groups. This classification was eventually achieved by the combined efforts of more than one hundred mathematicians in many published research articles and books written over a long period, the most intensive being 1955–80. It was a truly monumental feat of prolonged collaboration, and one of the most momentous theorems in the history of algebra.

In order to state the classification theorem, it is necessary to describe some examples of finite simple groups. The most obvious are the cyclic groups of prime order: these are clearly simple, since they have no subgroups at all apart from the identity and the whole group (by Lagrange’s theorem, for example, which states that the size of any subgroup is a factor of the size of the group). Next come the alternating groups A_n : here A_n is defined as the group consisting of all the even permutations in the symmetric group S_n (see PERMUTATION GROUPS [III.70]). The alternating group A_n has $\frac{1}{2}(n!)$ elements, and is simple provided $n \geq 5$. For example, A_5 , of order 60, is the smallest non-Abelian simple group.

Next we introduce some simple groups of matrices. For an integer $n \geq 2$ and a field K , define $SL_n(K)$ to be the set of all $n \times n$ matrices with entries in K and with DETERMINANT [III.15] equal to 1. This is a

group under matrix multiplication, called a *special linear* group. When the field K is finite, $SL_n(K)$ is a finite group. For each prime power q , there is up to isomorphism a unique field of order q , and the corresponding special linear group in dimension n is denoted by $SL_n(\mathbb{F}_q)$. These groups are not in general simple, since $Z = \{\lambda I : \lambda^n = 1\}$, the subgroup of scalar matrices in $SL_n(\mathbb{F}_q)$, is a normal subgroup. However, the factor groups $PSL_n(\mathbb{F}_q) = SL_n(\mathbb{F}_q)/Z$ are simple (except when $(n, q) = (2, 2)$ or $(2, 3)$). This is the family of *projective special linear* groups.

There are a number of other families of finite simple matrix groups, which, very roughly speaking, are defined as groups of matrices $A \in SL_n(\mathbb{F}_q)$ that satisfy an equation of the form $A^T J A = J$, where J is a non-singular symmetric or skew-symmetric $n \times n$ matrix. Again factoring out by the subgroup of scalar matrices, this gives the *projective orthogonal* and *symplectic* families of finite simple matrix groups. Similarly, if the finite field of order q has an automorphism $\alpha \rightarrow \bar{\alpha}$ of order 2, this can be extended to matrices $A = (a_{ij})$ by defining $\bar{A} = (\bar{a}_{ij})$, and then the group $\{A \in SL_n(\mathbb{F}_q) : A^T \bar{A} = I\}$, factored by its subgroup of scalar matrices, gives the *projective unitary* family of finite simple groups.

The families of projective special linear, symplectic, orthogonal, and unitary groups comprise what are known as the *classical* simple groups. These were all known early in the twentieth century, but it was not until 1955 that further infinite families of finite simple groups were discovered by Chevalley. For each of the simple complex Lie algebras L , and each finite field K , Chevalley constructed a version of L over K , call it $L(K)$, and defined his families of finite simple groups as automorphism groups of the Lie algebras $L(K)$. Not long afterward, Steinberg, Suzuki, and Ree found some variations of Chevalley’s construction and defined some further families of simple groups, known as twisted Chevalley groups. The Chevalley and twisted Chevalley groups include all the classical groups, together with ten other infinite families, and are collectively known as the *finite simple groups of Lie type*.

Until 1966, the only known finite simple groups were the cyclic groups of prime order, the alternating groups, the groups of Lie type, and a collection of five strange simple groups discovered by MATHIEU [VI.51] in the 1860s. These were groups of permutations of n objects, where $n = 11, 12, 22, 23$, or 24. Mathieu’s groups were termed “sporadic groups”—sporadic meaning that they do not fit into any of the

known infinite families—and many thought that perhaps there were no more finite simple groups to be found. Then there was a bombshell, when Janko published a paper demonstrating the existence of a single, new finite simple group: the sixth sporadic group. After this, new sporadic groups appeared at regular intervals, culminating in the MONSTER [III.63], an amazing group of order around 10^{54} , which was predicted by Fischer and constructed by Griess as a group of $196\,884 \times 196\,884$ matrices. By 1980, twenty-six sporadic groups were known.

During this period the program to classify all the finite simple groups was proceeding at breakneck speed, and eventually in the early 1980s the final classification theorem was announced.

Every finite simple group is either a cyclic group of prime order, or an alternating group, or a group of Lie type, or one of the twenty-six sporadic groups.

Not surprisingly, this theorem has changed the face of finite group theory and its many areas of application: one can now solve many problems in a concrete way, by reducing them to the study of the (now known) list of simple groups, rather than abstractly, by deducing them from the axioms for groups.

The sheer length of the proof of the classification theorem (estimated at around ten thousand journal pages, spread across about five hundred research articles) meant that it was extremely difficult, perhaps impossible, for a single person to work through the entire proof. It also meant that the chances were rather high that there were errors along the way. Fortunately, in the years since the announcement of the result, various teams of group theorists have been publishing summaries and revisions of many parts of the proof, and a series of volumes containing the whole proof is now well on the way to completion.

V.9 Dirichlet's Theorem

A famous theorem of EUCLID [VI.2] asserts that there are infinitely many primes. But what if one wants more information about these primes? For instance, are there infinitely many primes of the form $4n - 1$? A fairly straightforward modification of Euclid's argument shows that there are, and a slightly more difficult modification proves that there are infinitely many of the form $4n + 1$ as well. However, modifications of Euclid's argument are not enough to prove the general result in this direction, which is that if a and m are

coprime (that is, have highest common factor 1), then there are infinitely many primes of the form $mn + a$. This was proved by DIRICHLET [VI.36] using what are now called Dirichlet L -FUNCTIONS [III.49], which are closely related to the RIEMANN ZETA FUNCTION [IV.2 §3]. The condition that m and a have highest common factor 1 is clearly necessary, since any common factor of m and a will be a factor of $mn + a$. Dirichlet's theorem is discussed further in ANALYTIC NUMBER THEORY [IV.2 §4].

V.10 Dvoretzky's Theorem

Dvoretzky's theorem can be stated in two equivalent ways. On the one hand it is a central result in the theory of finite-dimensional BANACH SPACES [III.64] and on the other it is a highly counterintuitive geometrical statement about convex bodies.

The second formulation is easier to grasp and appreciate. A *convex body* is a shape in \mathbb{R}^n with the following property: given any two points in the shape, the line joining those two points lies entirely within the shape. Figure 1 shows two shapes, of which the first is convex and the second not convex (because the line joining the points A and B leaves the shape).

Suppose that we have chosen a point O to serve as an origin. Then, given any point P, there will be another point, usually called $-P$, at the same distance from O but in the opposite direction. A convex body is called *centrally symmetric* about O if, for every point P in the body, the opposite point $-P$ is also in the body. Thus, a square or circle is centrally symmetric while an equilateral triangle is not.

Figure 2 shows two cubes, which are examples of three-dimensional centrally symmetric convex bodies. Both cubes are cut by planes through their centers, and the parts of the planes that lie within the cubes are convex bodies themselves, but two-dimensional ones. They are known as *central cross sections* of the cubes. The first is a square and the second, more oblique one is a regular hexagon.

It is obvious that, whatever the angle of the plane that determines the cross section, the resulting two-dimensional body will not be a circle, since the faces of the cube are flat rather than curved. However, if we wanted to find the best approximation to a circle that we could, then we would do better to pick a regular hexagon than a square: the more sides a regular polygon has, the more circular it becomes. And this is in

PUP: Tim's answer to the proofreader's query here is as follows. "Whether or not something is easy to understand is not the same as whether it's intuitively plausible. In this case, the geometrical version is easier to grasp and this makes it easier to see just how counterintuitive it is." OK?

T&T note: position of figures in this article might be tricky. Check before page make-up stage.

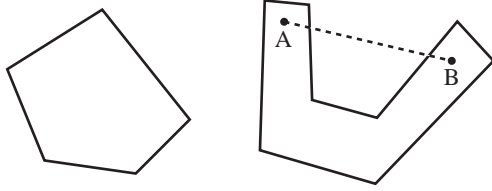


Figure 1 A convex and a nonconvex body.

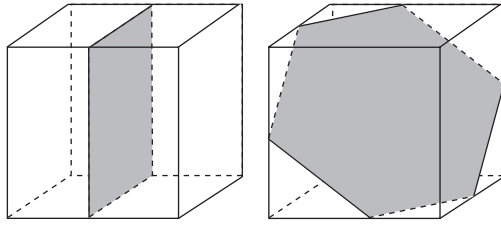


Figure 2 Two cross sections of a cube.

fact the best we can do, since no cross section of the cube has more than six sides.

However, remarkably, we *can* approximate a circle better if we look at cross sections of higher-dimensional bodies. There is a natural way to generalize all the above concepts to higher dimensions, and one can ask the following question. Let n and m be positive integers with m less than n and let K be an n -dimensional centrally symmetric convex body. Suppose we wish to find, among all the m -dimensional central cross sections of K , as good an approximation as we can to an m -dimensional sphere (which will be a circle when $m = 2$). How well can we do? Dvoretzky's theorem states that we can make the approximation as good as we like, provided only that n is chosen large enough. In other words, if you fix a dimension m and tell me how good you would like the approximation to be, then I can give you an n that will guarantee an approximation at least that good. To put this more loosely, a high-dimensional convex body has almost spherical cross sections, even if, like a cube, it has flat faces.

Dvoretzky proved his theorem in 1960. In 1970 Vitali Milman gave a second proof of Dvoretzky's theorem that was a triumph of the PROBABILISTIC METHOD [IV.19 §3]. His argument showed the following even more surprising fact: if the dimension of the convex body K is large enough, then *almost every* m -dimensional central cross section closely approximates an m -dimensional sphere. (Some care is needed in interpreting this statement.) The new proof led to intense activity in the theory of finite-dimensional

Banach spaces and there are now many extensions and modifications of the original theorem. For more discussion on convex geometry, see HIGH-DIMENSIONAL GEOMETRY AND ITS PROBABILISTIC ANALOGUES [IV.26].

V.11 Ergodic Theorems

Vitaly Bergelson

Consider the sequence $(z^n)_{n=0}^\infty$, where z is a complex number of modulus 1. While for $z \neq 1$ our sequence is not convergent, it is not hard to see that, on average, it exhibits quite regular behavior. Indeed, using the formula for the sum of a geometric progression, and assuming that $z \neq 1$, we have, for any $N > M \geq 0$,

$$\begin{aligned} \left| \frac{z^M + z^{M+1} + \dots + z^{N-1}}{N-M} \right| &= \left| \frac{z^M (z^{N-M} - 1)}{N-M} \right| \leq \frac{2}{(N-M)|z-1|}, \end{aligned}$$

which implies that when $N - M$ is large enough, the averages

$$A_{N,M}(z) = \frac{z^M + z^{M+1} + \dots + z^{N-1}}{N-M}$$

are small. More formally, we have

$$\lim_{N-M \rightarrow \infty} \frac{z^M + z^{M+1} + \dots + z^{N-1}}{N-M} = \begin{cases} 0, & z \neq 1, \\ 1, & z = 1. \end{cases} \quad (1)$$

This simple fact is a special, one-dimensional case of *von Neumann's ergodic theorem*, which was the first mathematical statement to throw light on the so-called quasi-ergodic hypothesis in statistical mechanics and the kinetic theory of gases.

Von Neumann's theorem concerns the average behavior of powers of UNITARY OPERATORS [III.52 §3.1] on HILBERT SPACES [III.37]. If U is such an operator defined on a Hilbert space \mathcal{H} , then we can associate with U the U -invariant subspace \mathcal{H}_{inv} that consists of all vectors $f \in \mathcal{H}$ such that $Uf = f$: that is, all vectors that are fixed by U . Let P be the ORTHOGONAL PROJECTION [III.52 §3.5] onto that subspace. Then von Neumann's theorem asserts that

$$\lim_{N-M \rightarrow \infty} \left\| \frac{1}{N-M} \sum_{n=M}^{N-1} U^n f - Pf \right\| = 0.$$

In other words, in a certain sense the averages

$$\frac{1}{N-M} \sum_{n=M}^{N-1} U^n$$

converge to the orthogonal projection P . (This is not actually the theorem as formulated by VON NEUMANN

[VI.91], but it is simpler to explain. He proved an equivalent statement about a continuous family of unitary operators $(U_\tau)_{\tau \in \mathbb{R}}$.

Before we discuss various applications and refinements of von Neumann's theorem, let us briefly comment on its proof. Von Neumann's original proof used sophisticated machinery such as the spectral theory of one-parameter groups of unitary operators, obtained by Marshall Stone. Over the years many alternative proofs were offered, the simplest being a "geometric" proof due to RIESZ [VI.74], which we will describe below. To give the rough idea of von Neumann's proof it is convenient to use the fact (which follows from the SPECTRAL THEOREM [III.52 §3.4]) that any unitary operator U on a Hilbert space \mathcal{H} has a "functional model." That is, we can realize the Hilbert space \mathcal{H} as a function space, consisting of all (equivalence classes of) square-integrable functions with respect to some finite MEASURE [III.57], in such a way that U becomes a *multiplication operator* $M_\varphi(f) = \varphi f$, where φ is a complex-valued measurable function that satisfies $|\varphi(x)| = 1$ for almost every x . It is not hard to see, after passing to such a functional model, that von Neumann's theorem follows immediately from its one-dimensional case as expressed by formula (1). Note that in this case the orthogonal projection to the space of invariant elements takes a function f to the function g such that $g(x) = f(x)$ if $\varphi(x) = 1$ and $g(x) = 0$ otherwise.

Riesz's proof is based on the observation that the orthogonal complement of the subspace \mathcal{H}_{inv} of U -invariant vectors is spanned by the set of vectors of the form $Ug - g$. To see this, note first that if $f \in \mathcal{H}_{\text{inv}}$, then

$$\langle f, Ug \rangle = \langle U^{-1}f, g \rangle = \langle f, g \rangle,$$

from which it follows that $\langle f, Ug - g \rangle = 0$ and thus that f is orthogonal to $Ug - g$. Conversely, if $f \notin \mathcal{H}_{\text{inv}}$, then $\langle f, Uf - f \rangle = \langle f, Uf \rangle - \langle f, f \rangle$. This is less than 0, by the CAUCHY-SCHWARZ INEQUALITY [V.22] and the fact that $\|Uf\| = \|f\|$ but $Uf \neq f$. In particular, f is not orthogonal to $Uf - f$. Thus, \mathcal{H}_{inv} is the orthogonal complement of the (closed) subspace of \mathcal{H} generated by functions of the form $Ug - g$.

Now the conclusion of von Neumann's theorem holds trivially if $f \in \mathcal{H}_{\text{inv}}$, since then $Pf = f$ and $U^n f = f$ for every n . On the other hand, if $f = Ug - g$, then $Pf = 0$. As for the averages, we know that $U^n f = U^{n+1}g - U^n g$, from which it follows that $\sum_{n=M}^{N-1} U^n f = U^N g - U^M g$. Since $\|U^N g - U^M g\|$ is at most $2\|g\|$ for every M and

N , we find that

$$\frac{1}{N-M} \sum_{n=M}^{N-1} U^n f$$

has norm at most $2\|g\|/(N-M)$ and hence tends to 0. So the theorem is true in this case as well. It is straightforward to check that the set of functions for which the theorem holds is a closed linear subspace of \mathcal{H} , and therefore the theorem is proved.

The reason that von Neumann's theorem and other similar results are relevant to physics is that it is often possible to represent the evolution of the parameters associated with a physical system by a subset $X \subset \mathbb{R}^d$ that has finite d -dimensional volume, together with a continuous family $(T_\tau)_{\tau \in \mathbb{R}}$ of volume-preserving transformations from X to X . With each such transformation T_τ one can associate the unitary map U_τ , defined on $L^2(X)$ (the Hilbert space of square-integrable functions on X) by the formula $(U_\tau f)(x) = f(T_\tau x)$. The fact that these maps are unitary follows from the fact that the transformations T_τ preserve volume; also, it follows from the fact that the transformations T_τ depend continuously on τ that the maps U_τ do as well.

To simplify the discussion let us now "discretize" the situation. Instead of considering the continuous families (T_τ) and (U_τ) we shall fix a transformation $T = T_{\tau_0}$ (say, for $\tau_0 = 1$) and let U be the corresponding unitary operator. Assume that our volume-preserving transformation T is *ergodic*, which means that there is no proper subset $A \subseteq X$ of positive volume such that $T(A) \subset A$. This assumption can easily be shown to be equivalent to the fact that the only elements of $L^2(X)$ that satisfy $Uf = f$ are the constant functions. It follows from von Neumann's theorem that for any $f \in L^2(X)$ the averages

$$A_{N,M}(f) = \frac{1}{N-M} \sum_{n=M}^{N-1} U^n f$$

converge to a constant whose value is easy to find by performing term-by-term integration: it equals $(\int f \, d\mu) / \text{vol}(X)$. Since von Neumann's theorem also tells us that $\lim_{N-M \rightarrow \infty} A_{N,M}(f)$ is always a U -invariant function, we see that the assumption of ergodicity is a necessary and sufficient condition for the time average represented by $\lim_{N-M \rightarrow \infty} A_{N,M}(f)$ to equal the space average, $(\int f \, d\mu) / \text{vol}(X)$.

One can also use von Neumann's theorem to strengthen a classical theorem of POINCARÉ [VI.61], called *Poincaré's recurrence theorem*. This result states that if X is a set of finite volume, as above, and A is

a subset of X with nonzero volume, then “almost all points of A return infinitely often to A .” In other words, if we set \tilde{A} to be the set of all points $x \in A$ such that $T^n x \in A$ for infinitely many n , then the measure of the set of points in A but not in \tilde{A} is 0. The main step in the proof of Poincaré’s theorem is to prove the same about the set A_1 , which consists of all points $x \in A$ such that $T^n x \in A$ for *some* positive integer n . To see why this is true, let B be the set of all points in A but not in A_1 . The sets $B, T^{-1}B, T^{-2}B, \dots$ all have the same measure, since T is volume preserving. ($T^{-n}B$ is defined to be the set of all x such that $T^n x \in B$.) Since X has finite volume, there must exist positive integers m and n such that the intersection of $T^{-m}B$ and $T^{-(m+n)}B$ has positive measure, and from this it follows that the measure of $B \cap T^{-n}B$ is also positive. But if $x \in B$ then $x \notin A_1$, so $T^n x \notin A$ and therefore $T^n x \notin B$, so this is a contradiction.

Now let us apply the von Neumann ergodic theorem with f equal to the characteristic function of a set A (that is, $f(x) = 1$ when $x \in A$ and $f(x) = 0$ otherwise) and U defined in terms of T as before. Suppose also that the set X has volume 1 and write μ for the measure on X . Then one can check that $\langle f, U^n f \rangle = \mu(A \cap T^{-n}A)$. It follows that

$$\langle f, A_{N,M}(f) \rangle = \frac{1}{N-M} \sum_{n=M}^{N-1} \mu(A \cap T^{-n}A).$$

If we let $N - M$ tend to infinity, then $A_{N,M}f$ tends to a U -invariant function g . Since g is U -invariant, $\langle f, g \rangle = \langle U^n f, g \rangle$ for every n , and therefore $\langle f, g \rangle = \langle A_{N,M}(f), g \rangle$ for every N and M , and finally $\langle f, g \rangle = \langle g, g \rangle$. By the Cauchy-Schwarz inequality, this is at least $(\int g(x) d\mu)^2 = (\int f(x) d\mu)^2 = \mu(A)^2$. Therefore, we deduce that

$$\lim_{N-M \rightarrow \infty} \frac{1}{N-M} \sum_{n=M}^{N-1} \mu(A \cap T^{-n}A) \geq (\mu(A))^2.$$

If you choose two “random sets” of measure $\mu(A)$, then their intersection will typically be $(\mu(A))^2$, so the inequality above is saying that the average intersection of A with $T^{-n}A$ is at least as big as the “expected” intersection. This result, due to Khinchin, gives more precise information about the nature of Poincaré recurrence.

When a unitary operator is defined in terms of a measure-preserving transformation as above, it is natural to ask whether the averages converge not just in the sense of the L^2 -norm but also in the more classical sense of convergence almost everywhere. (For a related thought in a different context, see CARLESON’S

THEOREM [V.5].) The answer is that they do, as was shown by BIRKHOFF [VI.78] soon after he learned of von Neumann’s theorem. He proved that for each integrable function f one could find a function f^* such that $f^*(Tx) = f^*(x)$ for almost every x , and such that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n x) = f^*(x)$$

for almost every x . Suppose that the transformation T is ergodic, let $A \subseteq X$ be a set of positive measure, and let $f(x)$ be the characteristic function of A . It follows from Birkhoff’s theorem that for almost every $x \in X$ one has

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n x) = \frac{\int f d\mu}{\mu(X)} = \frac{\mu(A)}{\mu(X)}.$$

Since the expression

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n x)$$

describes the frequency of visits of $T^n x$ to the set A , we see that in an ergodic system the images x, Tx, T^2x, \dots of a typical point $x \in A$ visit A with a frequency that equals the proportion of the space occupied by A .

The ergodic theorems of von Neumann and Birkhoff have been generalized over the years in many different directions. These far-reaching extensions of ergodic theorems, and more generally the *ergodic method*, have found impressive applications in such diverse fields as statistical mechanics, number theory, probability theory, harmonic analysis, and combinatorics.

Further Reading

- Furstenberg, H. 1981. *Recurrence in Ergodic Theory and Combinatorial Number Theory*. M. B. Porter Lectures. Princeton, NJ: Princeton University Press.
- Krengel, U. 1985. *Ergodic Theorems*, with a supplement by A. Brunel. De Gruyter Studies in Mathematics, volume 6. Berlin: Walter de Gruyter.
- Mackey, G. W. 1974. Ergodic theory and its significance for statistical mechanics and probability theory. *Advances in Mathematics* 12:178–268.

The Fermat-Euler Theorem

See MODULAR ARITHMETIC [III.60]

V.12 Fermat’s Last Theorem

Many people, even if they are not mathematicians, are aware of the existence of *Pythagorean triples*: that is,

triples of positive integers (x, y, z) such that $x^2 + y^2 = z^2$. These give us examples of right-angled triangles with integer side lengths, of which the best known is the “(3, 4, 5) triangle.” For any two integers m and n , we have that $(m^2 - n^2)^2 + (2mn)^2 = (m^2 + n^2)^2$, which gives us an infinite supply of Pythagorean triples, and in fact every Pythagorean triple is a multiple of a triple of this form.

FERMAT [VI.12] asked the very natural question of whether similar triples existed for higher powers: that is, could there be a solution in positive integers of the equation $x^n + y^n = z^n$ for some power $n \geq 3$? For instance, is it possible to express a cube as a sum of two other cubes? Or rather, Fermat famously claimed that it was not possible, and that he had a proof that space did not permit him to write down. Over the next three and a half centuries, this problem became the most famous unsolved problem in mathematics. Given the amount of effort that went into it, one can be virtually certain that Fermat did not in fact have a proof: the problem appears to be irreducibly difficult, and solvable only by techniques that were developed much later than Fermat.

The fact that Fermat's question was an easy one to think of does not on its own guarantee that it is interesting. Indeed, in 1816 GAUSS [VI.26] wrote in a letter that he found it too isolated a problem to interest him. At the time, that was a reasonable remark: it is often extremely hard to determine whether a given Diophantine equation has a solution, and it is therefore easy to come up with hard problems of a similar nature to Fermat's last theorem. However, Fermat's last theorem has turned out to be exceptional in ways that even Gauss could not have been expected to foresee, and nobody would now describe it as “isolated.”

By the time of Gauss's remark, the problem had been solved for $n = 3$ (by EULER [VI.19]) and $n = 4$ (by Fermat; this is the easiest case). The first serious connection between Fermat's last theorem and more general mathematical concerns came with the work of KUMMER [VI.40] in the middle of the nineteenth century. An important observation that had been made by Euler is that it can be fruitful to study Fermat's last theorem in larger RINGS [III.83 §1], since these, if appropriately chosen, allow one to factorize the polynomial $z^n - y^n$. Indeed, if we write $1, \zeta, \zeta^2, \dots, \zeta^{n-1}$ for the n th roots of 1, then we can factorize it as

$$(z - y)(z - \zeta y)(z - \zeta^2 y) \cdots (z - \zeta^{n-1} y). \quad (1)$$

Therefore, if $x^n + y^n = z^n$ then we have two rather different-looking factorizations of x^n inside the ring generated by 1 and ζ (namely the factorization in (1) above, and $xx \cdots x$), and it is reasonable to hope that this information might be exploited. However, there is a serious problem: the ring generated by 1 and ζ does not enjoy the UNIQUE FACTORIZATION PROPERTY [IV.1 §§4–8], so one's sense of being close to a contradiction when faced with these two factorizations is not well-founded. Kummer, in connection with the search for HIGHER RECIPROCITY LAWS [V.31], had met this difficulty and had defined the notion of an IDEAL [III.83 §2]: very roughly, if you enlarge a ring by adding in Kummer's “ideal numbers,” then unique factorization is restored. Using these concepts, Kummer was able to prove Fermat's last theorem for every prime number p that was not a factor of the CLASS NUMBER [IV.1 §7] of the corresponding ring. He called such primes *regular*. This connected Fermat's last theorem with ideas that have belonged to the mainstream of ALGEBRAIC NUMBER THEORY [IV.1] ever since. However, it did not solve the problem, since there are infinitely many irregular primes (though this was not known in Kummer's day).

It turned out that more complicated ideas could be used for individual irregular primes, and eventually an algorithm was developed that could check for any given n whether Fermat's last theorem was true for that n . By the late twentieth century, the theorem had been verified for all exponents up to 4 000 000. However, a general proof came from a very different direction.

The story of the eventual proof by Andrew Wiles has been told many times, so we shall be very brief about it here. Wiles did not study Fermat's last theorem directly, but instead solved an important special case of the *Shimura–Taniyama–Weil conjecture*, which connects ELLIPTIC CURVES [III.21] and MODULAR FORMS [III.61]. The first hint that elliptic curves might be relevant came when Yves Hellegouarch noticed that the elliptic curve $y^2 = x(x - a^p)(x - b^p)$ would have rather unusual properties if $a^p + b^p$ was also a p th power. Gerhard Frey realized that such a curve might be so unusual that it would contradict the Shimura–Taniyama–Weil conjecture. Jean-Pierre Serre came up with a precise statement (the “epsilon conjecture”) that would imply this, and Ken Ribet proved Serre's conjecture, thus establishing that Fermat's last theorem was a consequence of the Shimura–Taniyama–Weil conjecture. Wiles suddenly became very interested indeed, and after seven years of intensive and almost secret work he announced a solution to a case of the Shimura–

Taniyama–Weil conjecture that was sufficient to prove Fermat’s last theorem. It then emerged that Wiles’s proof contained a serious mistake, but with the help of Richard Taylor he managed to find an alternative and correct argument for that portion of the proof.

The Shimura–Taniyama–Weil conjecture asserts that “all elliptic curves are modular.” We finish by giving a rough idea of what this means. (A few more details can be found in ARITHMETIC GEOMETRY [IV.5].) Associated with any elliptic curve E is a sequence of numbers $a_n(E)$, one for each positive integer n . For each prime p , $a_p(E)$ is related to the number of points on the elliptic curve (mod p); it is easy to derive from these values the values of $a_n(E)$ for composite n . Modular forms are HOLOMORPHIC FUNCTIONS [I.3 §5.6] with certain periodicity properties defined on the upper half-plane; associated with each modular form f is a FOURIER SERIES [III.27] that takes the form

$$f(q) = a_1(f)q + a_2(f)q^2 + a_3(f)q^3 + \cdots$$

Let us call an elliptic curve E *modular* if there is a modular form f such that $a_p(E) = a_p(f)$ for all but finitely many primes p . If you are presented with an elliptic curve, it is not at all clear how to set about finding a modular form associated with it in this way. However, it always seemed to be possible, even if the phenomenon was a mysterious one. For instance, if E is the elliptic curve $y^2 + y = x^3 - x^2 - 10x - 20$, then there is a modular form f such that $a_p(E) = a_p(f)$ for every prime p apart from 11. This modular form is the unique complex function (up to scaling) that satisfies a certain periodicity property with respect to the group $\Gamma_0(11)$, which consists of all matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ such that a, b, c , and d are integers, c is a multiple of 11, and the DETERMINANT [III.15] $ad - bc$ is 1. It is far from obvious that a definition of this type should have anything to do with elliptic curves.

Wiles proved that all “semistable” elliptic curves are modular, not by showing how to associate a modular form with each such elliptic curve, but by using a subtle counting argument that guaranteed that the modular form had to exist. The full conjecture was proved a few years later, by Christophe Breuil, Brian Conrad, Fred Diamond, and Richard Taylor, which put the icing on the cake of one of the most celebrated mathematical achievements of all time.

V.13 Fixed-Point Theorems

1 Introduction

The following is a variant of a well-known mathematical puzzle. A man is on a train from London to Cambridge and has a bottle of water with him. Prove that there is at least one moment on the journey when the volume of air in the bottle, as a fraction of the volume of the bottle itself, is exactly equal to the fraction of his journey that he has completed. (For instance, the bottle might be two fifths full, and therefore three fifths empty, at the precise moment when he is three fifths of the way from London to Cambridge. Note that we do not assume that the bottle is full at the start of the journey or empty at the end.)

The solution, if you have not seen this sort of question before, is surprisingly simple. For each x between 0 and 1 let $f(x)$ be the proportion of air in the bottle when the proportion of the journey that has been completed is x . Then $0 \leq f(x) \leq 1$ for every x , since the volume of air in the bottle cannot be negative and cannot exceed the volume of the bottle. If we now set $g(x)$ to be $x - f(x)$, then we see that $g(0) \leq 0$ and $g(1) \geq 0$. Since $g(x)$ varies continuously with x , there must be some moment at which $g(x) = 0$, so that $f(x) = x$, which is what we wanted.

What we have just proved is a slightly disguised form of one of the simplest of all fixed-point theorems. We could state it more formally as follows: if f is a continuous function from the closed interval $[0, 1]$ to itself, then there must exist an x such that $f(x) = x$. This x we call a *fixed point* of f . (We deduced the result from the *intermediate-value theorem*, a basic result in analysis that states that if g is a continuous function from $[0, 1]$ to \mathbb{R} such that $g(0) \leq 0$ and $g(1) \geq 0$, then there must be some x such that $g(x) = 0$.)

In general, a fixed-point theorem is a theorem that asserts that a function that satisfies certain conditions must have a fixed point. There are many such theorems, a small sample of which we shall discuss in this article. On the whole, they tend to have a nonconstructive nature: they establish the existence of a fixed point rather than defining one or telling you how to find it. This is part of the reason that they are important, since there are many examples of equations for which one would like to prove that a solution exists even when one cannot solve it explicitly. As we shall see, one way of going about this is to try to rewrite the equation in the form $f(x) = x$ and apply a fixed-point theorem.

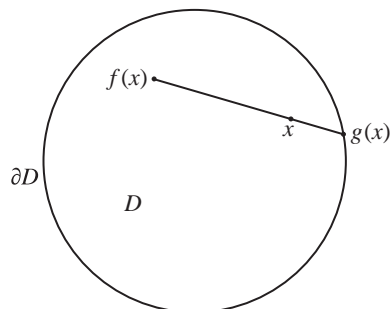


Figure 1 If f has no fixed points, then it can be used to define a retraction g .

2 Brouwer's Fixed-Point Theorem

The fixed-point theorem we have just proved is the one-dimensional version of *Brouwer's fixed-point theorem*, which states that if B_n is the unit ball of \mathbb{R}^n (that is, the set of all (x_1, \dots, x_n) such that $x_1^2 + \dots + x_n^2 \leq 1$) and f is a continuous function from B_n to B_n , then f must have a fixed point. The set B_n is an n -dimensional solid sphere, but all that matters is its topological character, so we could take it to be another shape such as an n -dimensional cube or simplex.

In two dimensions this says that a continuous function from the closed unit disk to itself must have a fixed point. In other words, if you had a circular sheet of rubber on a table and you picked it up and put it back down within the circle where it started, having folded it and stretched it as much as you liked, there would always have to be a point that ended up in the same place as before.

To see why this is true, it is helpful to reformulate the statement. Let $D = B_2$ be the closed unit disk. If we had a continuous function f from D to D with no fixed point, then we could define a continuous function g from D to its boundary ∂D as follows: for each x , follow a straight path from $f(x)$ to x and continue on in a straight line; $g(x)$ is the point where you first reach ∂D (see figure 1), and it is well-defined because (and only because) $f(x) \neq x$. If x is already on the boundary of D , then $g(x) = x$. So we have a continuous function $g : D \rightarrow \partial D$ such that $g(x) = x$ for every $x \in \partial D$. Such a function is called a *retraction* from D to ∂D .

It seems highly unlikely that a continuous retraction from D to ∂D could exist. If we can prove that it cannot, then we will have contradicted the assumption that there is a continuous function from D to D with no fixed point, and thereby have proved Brouwer's fixed-point theorem in two dimensions.

There are several ways of proving that continuous retractions from disks to their boundaries cannot exist. Here we briefly sketch two.

Suppose, first, that g is such a retraction. For each t , let us consider the restriction of g to the circle of radius t about the origin, and let us represent a typical point in this circle as $te^{i\theta}$. Let us write $g_t(\theta)$ for $g(te^{i\theta})$. When $t = 1$ the circle of radius t is ∂D , so as θ goes from 0 to 2π , $g_t(\theta) = e^{i\theta}$ goes once around the unit circle. When $t = 0$, the circle of radius t is a single point, so as θ goes from 0 to 2π , $g_t(\theta)$ is just the constant point $g(0)$, which does not go around the unit circle at all. Therefore, somewhere between $t = 1$ and $t = 0$ there must be a change in the number of times $g_t(\theta)$ goes around the unit circle as θ goes from 0 to 2π . But the functions g_t are a continuously varying family of functions, and a small change in g_t cannot cause a sudden jump in the number of times that $g_t(\theta)$ goes around the circle. (To make this last step rigorous needs a bit of work, but the basic idea is sound.)

A second proof uses basic tools from algebraic topology. The first HOMOLOGY GROUP [IV.6 §4] of the disk D is trivial, since every curve in the disk can be shrunk to a point. The first homology group of the unit circle ∂D is \mathbb{Z} . If there is a continuous retraction g from D to ∂D , then we can find continuous maps $h : \partial D \rightarrow D$ and $g : D \rightarrow \partial D$ such that $g \circ h$ is the identity on ∂D . (We let h be the map that takes a point of ∂D to itself and we let g be the continuous retraction.) Now continuous maps between topological spaces give rise to HOMO-MORPHISMS [I.3 §4.1] between their homology groups, in such a way that compositions go to compositions and identity maps go to identity maps. (That is, there is a FUNCTOR [III.8] from the CATEGORY [III.8] of topological spaces and continuous maps to the category of groups and group homomorphisms.) This means that there must be homomorphisms $\phi : \mathbb{Z} \rightarrow \{0\}$ and $\psi : \{0\} \rightarrow \mathbb{Z}$ such that $\psi \circ \phi$ is the identity on \mathbb{Z} , which is obviously impossible.

Both proofs generalize to higher dimensions: the second straightforwardly (once one knows how to compute homology groups of spheres), and the first via the notion of the *degree* of a continuous map from the n -sphere to itself, which is a higher-dimensional analogue of the notion of the number of times a map from the circle to itself "goes around the circle."

Brouwer's fixed-point theorem has many applications. For example, the following fact is important in the theory of random walks on graphs. A *stochastic matrix* is an $n \times n$ matrix with nonnegative entries such

that the sum of the entries in each row is equal to 1. Brouwer's fixed-point theorem can be used to show that every such matrix has an EIGENVECTOR [I.3 §4.3] with nonnegative entries and eigenvalue 1. The proof is as follows: the set of all column vectors with nonnegative entries that add up to 1 is, geometrically speaking, an $(n - 1)$ -dimensional simplex. (For example, if $n = 3$, this set is a triangle in \mathbb{R}^3 with vertices $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$.) If A is a stochastic matrix and \mathbf{x} belongs to this simplex, then so does $A\mathbf{x}$. Since the map $\mathbf{x} \mapsto A\mathbf{x}$ is continuous, Brouwer's theorem gives us an \mathbf{x} such that $A\mathbf{x} = \mathbf{x}$: this is the required eigenvector.

An extension of Brouwer's theorem, called the *Kakutani fixed-point theorem*, was used by John Nash to establish the existence of a "social equilibrium," a state of affairs in which no household can individually improve its well-being by altering the amount that it consumes of various items. Kakutani's theorem concerns functions that take points in a closed ball B not to other points in B but to *subsets* of B . If $f(x)$ is a nonempty closed convex subset of B for each x and if $f(x)$ varies continuously in an appropriate sense, then the theorem says that there must be some x such that $x \in f(x)$. Brouwer's theorem is the special case where each $f(x)$ is a set with just one element.

3 A Stronger Form of Brouwer's Fixed-Point Theorem

So far, we have discussed maps from solid spheres to themselves, but there is nothing to stop us thinking about whether continuous maps on other spaces must have fixed points. For example, let S^2 be the (nonsolid) sphere $\{(x, y, z) : x^2 + y^2 + z^2 = 1\}$ and let f be a continuous function from S^2 to S^2 . Must f have a fixed point? At first one might think so: some obvious functions from S^2 to itself are rotations and reflections, both of which certainly have fixed points, and it is hard to see how one can "get rid" of those fixed points. However, eventually one realizes that there is a simple example of a function without a fixed point, namely the function $f(x) = -x$, which reflects each point through the origin.

The obvious reaction to this example is to note that the result we had hoped for is false and to turn our attention to something else. But this reaction is a mistake, as it is in many other mathematical contexts, because there was something importantly correct about the idea that it was impossible to get rid of the fixed points of a rotation. It turns out that if you start

with a rotation and try to get rid of the fixed points by continuously deforming it, then you are doomed to failure. In fact, in a certain sense there will always be exactly two fixed points. More generally, if you take any continuous function from S^2 to S^2 and continuously deform it, then you cannot change the number of fixed points.

Of course, these last two statements are patently false if taken at face value so some reinterpretation is needed. First, we must assume that the number of fixed points is finite, but this is not a huge assumption as it can be shown that a typical small perturbation of any continuous function will have only finitely many fixed points. Second, we must count the fixed points with appropriate weights. To define these, suppose that $f(x) = x$, and imagine a point $y(t)$ that goes around x in a tiny circle as t goes from 0 to 1. We define the *index* of the fixed point x to be the number of times that $f(y(t))$ goes around x , counting this negatively if it goes around in the opposite direction to $y(t)$. (This definition is problematic if $f(y(t)) = x$ for some t , but again we can make small perturbations and assume that this does not happen.) Then the sum of the indices of all the fixed points is the quantity that does not change if you continuously deform f .

It follows that if you continuously deform a rotation, then the sum of the indices will always be 2. From this it follows that there must be at least one fixed point. It also follows that you cannot continuously deform a rotation so that it becomes the map that sends each x to $-x$.

The notion of the index of a fixed point can be generalized in a fairly straightforward way to higher dimensions (using the concept of degree mentioned earlier), and one can show under very general circumstances that the sum of the indices of fixed points remains constant when you continuously deform a continuous map. This implies Brouwer's fixed-point theorem as follows. We can continuously deform any continuous map $f : B_n \rightarrow B_n$ into to any other continuous map $g : B_n \rightarrow B_n$ by defining $f_t(x) = (1 - t)f(x) + tg(x)$ and letting t vary from 0 to 1. Let us therefore take g to be the map $x \mapsto \frac{1}{2}x$, which has a single fixed point. This fixed point has index 1 (as one can see easily in the two-dimensional case), and therefore the sum of the indices of the fixed points of f is 1 as well.

In general, the sum of the indices of the fixed points of a function f defined on a suitable topological space X (such as a smooth compact MANIFOLD [I.3 §6.9]) can

be calculated in terms of the effect of f on the homology groups of X . The resulting theorem is (a slight generalization of) the *Lefschetz fixed-point theorem*.

The fact that the index of a continuous map is an invariant of continuous deformations can be used to give a proof of THE FUNDAMENTAL THEOREM OF ALGEBRA [V.15]. Consider, for instance, the problem of proving that the polynomial $x^5 + 3x + 8$ has a root. This is the same as asking for a fixed point of the function $x^5 + 4x + 8$, since if this equals x then $x^5 + 3x + 8 = 0$. Now if we regard the polynomial x^5 as being defined on the RIEMANN SPHERE [IV.14 §2.4] $\mathbb{C} \cup \{\infty\}$, then it has two fixed points, at 0 and ∞ . Moreover, their indices are both 5 (since if x goes around 0 or ∞ in a “small circle,” then x^5 goes around five times). Now the polynomials $x^5 + (4x + 8)t$ give us a continuous deformation from x^5 to $x^5 + 4x + 8$, and $x^5 + 4x + 8$ has a fixed point of index 5 at ∞ . It follows that there must be other fixed points, with indices adding up to 5. These are the roots of $x^5 + 3x + 8$, and the indices are the multiplicities of the roots.

4 Infinite-Dimensional Fixed-Point Theorems and Applications to Analysis

What happens if we try to generalize the Brouwer fixed-point theorem to continuous maps defined on infinite-dimensional closed balls? The answer is that we will not be able to, as the following example shows. Let B be the set of all sequences (a_1, a_2, \dots) such that $\sum_n |a_n|^2 \leq 1$. This is our closed ball; it is the unit ball of the HILBERT SPACE [III.37] ℓ_2 . Given an infinite sequence $\mathbf{a} = (a_1, a_2, \dots)$, we write $\|\mathbf{a}\|$ for its norm $(\sum_n |a_n|^2)^{1/2}$. Now consider the map $f: (a_1, a_2, \dots) \mapsto ((1 - \|\mathbf{a}\|^2)^{1/2}, a_1, a_2, \dots)$. It is easy to check that f is continuous and that $\|f(\mathbf{a})\| = 1$ for every \mathbf{a} . Therefore, if \mathbf{a} is a fixed point, we must have $\|\mathbf{a}\| = 1$, from which we can see that $a_1 = 0$. From this it follows that $a_2 = 0$, and then that $a_3 = 0$, and so on. In other words, $\mathbf{a} = 0$. But this contradicts the condition that $\|\mathbf{a}\| = 1$. Therefore, the map f has no fixed point.

However, if we place extra conditions on a continuous map, then it is sometimes possible to prove fixed-point theorems, and some of these theorems have important applications, notably to establishing the existence of solutions to differential equations.

An easy result of this type is the *contraction mapping theorem*. This states that if X is a METRIC SPACE [III.58] with a property known as *completeness* (which is briefly discussed in NORMED SPACES AND BANACH

SPACES [III.64]) and f is a map from X to X such that there exists a constant $\rho < 1$ such that $d(f(x), f(y)) \leq \rho d(x, y)$ for every x and y in X , then f must have a fixed point. To prove this, one picks any point $x \in X$ and looks at the iterates $x, f(x), f(f(x)), f(f(f(x)))$, and so on. Denoting these by x_0, x_1, x_2, \dots , one can prove quite easily that $d(x_n, x_m)$ tends to 0 as m and n both tend to infinity, and the completeness property then guarantees that the sequence (x_n) has a limit. It is not hard to prove that this limit is a fixed point of f .

A more sophisticated example is the *Schauder fixed-point theorem*, which states that if X is a Banach space, K is a COMPACT [III.9] convex subset of X , and f is a continuous function from K to K , then f has a fixed point. Roughly speaking, to prove this one approximates K by larger and larger finite-dimensional sets K_n and approximates f by continuous maps f_n that take K_n to K_n . Brouwer's fixed-point theorem gives a sequence (x_n) such that $f_n(x_n) = x_n$ for each n . The compactness of K implies that the sequence (x_n) has a convergent subsequence: its limit can be shown to be a fixed point of f .

The importance of these two theorems, and others of a similar nature, lies more in their applications than in their basic statements. A typical application is a proof that the differential equation

$$\frac{d^2 u}{dx^2} = u - 10 \sin(u^2) - 10 \exp(-|x|)$$

has a solution u such that $u(x)$ is defined for every real number x and tends to 0 as x tends to $\pm\infty$. We can rewrite this equation as

$$\left(1 - \frac{d^2}{dx^2}\right)u = 10 \sin(u^2) + 10 \exp(-|x|).$$

If we write the left-hand side as $L(u)$, then this equation can be further rewritten as

$$u = L^{-1}(10 \sin(u^2) + 10 \exp(-|x|)).$$

(It is possible to identify the operator L^{-1} explicitly.) If we now let X be the Banach space of continuous functions defined on \mathbb{R} that tend to 0 at $\pm\infty$, with the uniform norm, then it can be shown that the right-hand side of this last equation defines a continuous function from X to a compact convex subset of X . Therefore, by the Schauder fixed-point theorem, this highly nonlinear equation has a solution with the given boundary conditions, a result that is hard to prove in any other way.

PUP: Tim says that any momentary puzzlement here is probably a good thing, so would like to keep it as it is. OK?

V.14 The Four-Color Theorem

Bojan Mohar

The four-color theorem asserts that the regions of any map drawn in the plane (or, equivalently, on the two-dimensional sphere) can be colored with no more than four colors in such a way that any two regions with a common boundary are given different colors. The example in figure 1 shows that four distinct colors are necessary since the regions A, B, C, and D are all adjacent to each other. This result was conjectured by Francis Guthrie in 1852. An incorrect proof was given by Kempe in 1879, and for eleven years the problem was believed to have been solved, until Heawood pointed out the error in 1890. However, Heawood showed that Kempe's basic idea, which we shall outline below, could at least be used to give a correct proof that five colors were always sufficient. After that, the problem became a famous example of a question that remained stubbornly open despite being very easy to understand. (Another such problem was FERMAT'S LAST THEOREM [V.12].)

In modern mathematics, map-coloring problems are usually formulated in the language of graph theory. To any map we assign a GRAPH [III.34]: the vertices of the graph correspond to the regions of the map, and we declare two vertices to be adjacent if the corresponding regions share a piece of their boundary. The graph for the map in figure 1 is shown in figure 2. It is easy to see that the graph of any map in the plane can be drawn in such a way that no two edges cross each other: such graphs are called *planar*. Instead of coloring regions of maps, we now color vertices of the corresponding graphs. If no two vertices that are joined by an edge have the same color, then we say that the coloring is *proper*. After this reformulation, the four-color theorem states that every planar graph G has a proper coloring with at most four colors.

Here, briefly, is the proof of the *five-color* theorem due to Kempe and Heawood. It is a proof by contradiction, so we start by assuming that the result is false. If that is the case, then there must be a graph G of minimal size that has no proper coloring with five colors. EULER'S FORMULA [I.4 §2.2] says that $V - E + F = 2$ for any (connected) planar graph, where V is the number of vertices, E is the number of edges, and F is the number of regions into which the plane is divided by any drawing of the graph. It is not hard to deduce from this

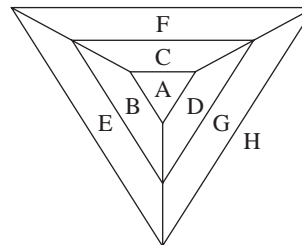


Figure 1 A map with eight regions.

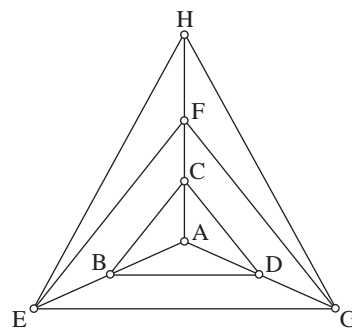


Figure 2 The graph of the map from figure 1.

formula that G has a vertex v with at most five neighbors (that is, other vertices linked to v by an edge) in the graph. If we remove v from the graph, then we can find a proper coloring of what is left, because G is a minimal counterexample to the theorem. If v has *fewer* than five neighbors, then we can color v as well, since there are at most four colors that need to be avoided and we have five colors at our disposal. So the only thing that can go wrong is if v has five neighbors and those five colors all get different colors when we color the rest of G .

Let us suppose that the colors of the neighbors of v are red, yellow, green, blue, and brown, as we go clockwise around v . As it stands, we cannot color v , but we could try to do so by adjusting the coloring of the rest of the graph. For instance, we could try recoloring the red vertex green, thereby freeing up red to be used for v . Of course, if we did that we might have to recolor further vertices, but we could try to find a recoloring as follows: first change the color of the red neighbor of v to green. Then change all the green neighbors of *that* vertex to red, and all the red neighbors of those vertices to green, and so on. When we have finished this process, the one thing that could go wrong is that we might end up recoloring the green neighbor of v red, in which case we would not after all be free to use red for

v . This will happen if and only if there is a chain of vertices from the red neighbor of v to the green neighbor that alternates red and green. However, if this circumstance arises, we can try to recolor the yellow neighbor of v blue in a similar way. Once again, the only thing that can stop us is an alternating chain of yellow and blue vertices going from the yellow neighbor of v to the blue neighbor. But such a chain cannot exist, as it would at some point have to cross the red/green chain, and this contradicts the fact that the graph is planar.

Returning to the four-color problem, the German mathematician Heinrich Heesch proposed a general method for tackling it that can be thought of as a more complicated version of the above argument. The idea is to identify a list C of “configurations” with the following properties. First, every planar graph must contain a configuration X that belongs to C . Second, given a planar graph G that contains a configuration X from C , and given a proper coloring of the rest of G that uses at most four colors, it is possible to adjust this coloring in such a way that it can be extended to a proper coloring of the whole of G . In the proof of the five-color theorem above, there was a very simple list of five configurations: a vertex v with one edge, two edges, three edges, four edges, or five edges coming out of it. Nothing this simple works for the four-color problem, but Heesch’s idea was that it might be possible to solve the problem by using a more complicated list of configurations.

Such a list was found by Kenneth Appel and Wolfgang Haken in 1976. However, this is by no means the whole story, because the list of configurations that they found was not just “more complicated” but so much more complicated that it broke new ground: it was the first time that a major theorem had been proved with a proof that was too long to be humanly checkable. The reason for this was partly that their list C contained about 1200 configurations, but a more important reason was that for some configurations X it was necessary to check hundreds of thousands of cases in order to demonstrate that a coloring of the rest of the graph could be adjusted to accommodate a coloring of X as well. Therefore, there was no alternative but to use a computer to do the checking. (Heesch had himself proposed a list, but some of his configurations would have involved so many cases that even a computer could not have checked them all.)

The reaction of other mathematicians to the proof of Appel and Haken was mixed. Some hailed it as the addition of a powerful new tool to the mathematical armory.

Others were uneasy about having to trust that the relevant computer program had been written correctly and that the computer had operated as it should. And in fact the proof turned out to have several flaws, though all those that were discovered were subsequently corrected by Appel and Haken in their monograph of 1989. Any doubts there may have been of this kind were removed once and for all in 1997, when Robertson, Sanders, Seymour, and Thomas developed another proof based on similar principles. The part of the proof that was checkable by humans was made more transparent, and the computer-verified part was supported by a well-structured collection of data that enabled the proofs to be checked independently. One could still question whether the compilers used were correct and whether the hardware was stable, but the proof has been checked on different platforms, using different programming languages and operating systems, so this proof is much less likely to be incorrect than a typical human-checked proof of even moderate length.

The result is that very few mathematicians are now worried about whether the proof is correct. However, there are many who object to it for a different reason. Even if we can now be certain that the theorem is true, we can still ask *why* it is true, and not everybody regards the answer “Because hundreds of thousands of cases were checked and they all turned out to be OK” as a satisfactory explanation. As a result, if someone were to discover a shorter and more accessible proof it would be regarded by many as a breakthrough comparable to the solution of the problem by Appel and Haken. An unfortunate side effect of this is that mathematics departments around the world still receive many incorrect attempted proofs, several of which repeat the mistake of Kempe.

Like many good problems, the four-color problem provoked the development of many important new mathematical ideas. The theory of graph colorings, in particular, has evolved into a deep and beautiful area of research. (See EXTREMAL AND PROBABILISTIC COMBINATORICS [IV.19 §2.1.1] and also Jensen and Toft (1995).) Extensions of map-coloring problems to arbitrary surfaces led to the development of topological graph theory, and questions about the planarity of graphs culminated in the theory of GRAPH MINORS [V.35].

One of the most prolific graph theorists, William T. Tutte, judged the impact of the four-color theorem on mathematics by proclaiming: “The four-colour theorem is the tip of the iceberg, the thin end of the wedge, and the first cuckoo of Spring.”

Further Reading

- Appel, K., and W. Haken. 1976. Every planar map is four colorable. *Bulletin of the American Mathematical Society* 82:711–12.
- . 1989. *Every Planar Map Is Four Colorable*. Contemporary Mathematics, volume 98. Providence, RI: American Mathematical Society.
- Jensen, T., and B. Toft. 1995. *Graph Coloring Problems*. New York: John Wiley.
- Robertson, N., D. Sanders, P. Seymour, and R. Thomas. 1997. The four-colour theorem. *Journal of Combinatorial Theory B* 70:2–44.

V.15 The Fundamental Theorem of Algebra

The COMPLEX NUMBERS [I.3 §1.5] can be thought of as what you obtain from the REAL NUMBERS [I.3 §1.4] when you introduce a new number, denoted i , and stipulate that it is a solution of the equation $x^2 = -1$, or equivalently a root of the polynomial $x^2 + 1$. At first, this may seem an artificial thing to do—it is not obvious what is so important about $x^2 + 1$ as opposed to any other polynomial—but that is a judgment with which no professional mathematician would concur. The fundamental theorem of algebra is one of the best pieces of evidence that the complex number system is, in fact, natural, and natural in a profound way. It states that, within the complex number system, *every* polynomial has a root. In other words, once we introduce the number i , then not only can we solve the equation $x^2 + 1 = 0$, we can solve all polynomial equations (even if the coefficients are themselves complex). Thus, when one defines the complex numbers, one gets much more out of them than one puts in. It is this that makes them seem not an artificial construction but a wonderful discovery.

For many polynomials it is not hard to see that they have roots. For example, if $P(x) = x^d - u$ for some positive integer d and some complex number u , then a root of P will be a d th root of u . One can write u in the form $re^{i\theta}$, and then $r^{1/d}e^{i\theta/d}$ will be such a root. This means that any polynomial that can be solved by a formula involving d th roots and the usual arithmetical operations, which includes all polynomials of degree less than 5, can be solved in the complex number system. However, owing to THE INSOLUBILITY OF THE QUINTIC [V.24], not all polynomials can be dealt with in this way, and in order to prove the fundamental theorem of algebra one must look for a less direct argument.

In fact, this is true even if one is looking for real roots of real polynomials. For example, if $P(x) = 3x^7 - 10x^6 + x^3 + 1$, then we know that $P(x)$ is large and positive when x is, since the x^7 term is by far the most significant, and large and negative when x is, for the same reason. Therefore, at some point the graph of P crosses the x -axis, which means that there is some x with $P(x) = 0$. Notice that this argument does not tell us what x is—that is the sense in which it is “less direct.”

Now let us see how one might show that a polynomial has a complex root, by looking at the example $P(x) = x^4 + x^2 - 6x + 9$. This can be rewritten $x^4 + (x - 3)^2$, and since both x^4 and $(x - 3)^2$ are nonnegative, and since they cannot be zero simultaneously, P cannot have a real root. To see that it has a complex root, we shall begin by fixing a large real number r and looking at the behavior of $P(re^{i\theta})$ as θ varies between 0 and 2π . As θ varies in this way, $re^{i\theta}$ traces out a circle of radius r in the complex plane.

Now $(re^{i\theta})^4 = r^4e^{4i\theta}$, so the x^4 part of $P(re^{i\theta})$ traces out a circle of radius r^4 , but goes around it four times. If r is large enough, then the rest (that is, $(re^{i\theta} - 3)^2$) is so small compared with $(re^{i\theta})^4$ that the only effect on the behavior of $P(re^{i\theta})$ is to make it deviate very slightly from the circle of radius r^4 . This small deviation is not enough to stop the path of $P(re^{i\theta})$ going around zero four times.

Next, let us consider what happens when r is very small. Then $P(re^{i\theta})$ is very close to 9, whatever the value of θ , since $(re^{i\theta})^4$, $(re^{i\theta})^2$, and $(re^{i\theta})$ are all small. But this means that the path traced out by $P(re^{i\theta})$ does not go around zero at all.

For any r we can ask how many times the path traced out by $P(re^{i\theta})$ goes around zero. What we have just established is that for very large r the answer is four and for very small r it is zero. It follows that at some intermediate r the answer changes. But if you gradually shrink r , the path traced out by $P(re^{i\theta})$ varies in a continuous way, so the only way this change can come about is if for some r the path crosses 0. This gives us the root we are looking for, since the path consists of points of the form $P(re^{i\theta})$ and one of these points is 0.

Some care is needed to turn the above reasoning into a rigorous proof. However, this can be done, and it is not hard to generalize the resulting argument to one that applies to any polynomial.

The fundamental theorem of algebra is usually attributed to GAUSS [VI.26], who proved it in 1799 in his doctoral thesis. Though his argument (which was

different from the one sketched above) was not fully rigorous by today's standards, it was convincing and broadly correct. Later he went on to give three more proofs.

V.16 The Fundamental Theorem of Arithmetic

The fundamental theorem of arithmetic is the assertion that every positive integer can be expressed in exactly one way as a product of prime numbers. These prime numbers are known as the *prime factors* of the original number and the product itself is the *prime factorization*. To give a few examples: $12 = 2 \times 2 \times 3$, $343 = 7 \times 7 \times 7$, $4559 = 47 \times 97$, and 7187 is itself a prime. This last number shows that the word “product” should be interpreted so as to include the case where there is only one prime involved. As for the phrase “exactly one way,” it is understood that the order in which the primes are multiplied is not significant, so, for example, the products 47×97 and 97×47 are not regarded as different.

The following inductive procedure allows one to find the prime factorization of a given positive integer n . If n is prime, then we have found it already. Otherwise, let p be the smallest prime factor of n and let $m = n/p$. Since m is smaller than n , we know by induction how to find the prime factorization of m , and this, together with p , gives it to us for n . In practice, what this means is that we generate a sequence of numbers, where each number in the sequence is the previous one divided by its smallest prime factor. For example, if we start with the number 168, then the sequence begins 168, 84, 42, 21. At this point we cannot divide by 2, but 3 is a factor of 21 so the next number in the sequence is 7. Since 7 is a prime, the process stops. Looking back, we find that we have shown that $168 = 2 \times 2 \times 2 \times 3 \times 7$.

Once one is used to this method, it comes to seem inconceivable that a number could have two genuinely different prime factorizations. But the method does not guarantee this at all. Suppose we successively divide by the largest prime factor rather than the smallest. Why should this not give a completely different set of primes? It is hard to think of an argument that does not use a phrase such as “the prime factorization of n ,” thereby implicitly assuming what it sets out to prove.

It is possible to show in a rather precise way that the fundamental theorem of arithmetic is *not* obvious, by looking at an algebraic structure where the notion of prime factorization makes sense but numbers can

have more than one prime factorization. This structure, denoted $\mathbb{Z}(\sqrt{-5})$, is the set of all numbers of the form $a + b\sqrt{-5}$, where a and b are integers. Such numbers can be added and multiplied just like ordinary integers. For example,

$$(1 + 3\sqrt{-5}) + (6 - 7\sqrt{-5}) = 7 - 4\sqrt{-5}$$

and

$$\begin{aligned} (1 + 3\sqrt{-5})(6 - 7\sqrt{-5}) &= 6 - 7\sqrt{-5} + 18\sqrt{-5} + 28(\sqrt{-5})^2 \\ &= 6 + 11\sqrt{-5} - 28 \times 5 \\ &= -134 + 11\sqrt{-5}. \end{aligned}$$

In this structure, we can regard a number $x = a + b\sqrt{-5}$ as prime if its only factors are ± 1 and $\pm x$. (This would also be a natural definition if we wanted to extend the notion of primes from the positive integers to all integers.) It can be shown quite easily that 2 and 3 are both primes (though it is not immediately obvious since there are now more possibilities for factors). Two other primes are $1 + \sqrt{-5}$ and $1 - \sqrt{-5}$. But we can write 6 either as 2×3 or as $(1 + \sqrt{-5})(1 - \sqrt{-5})$, so 6 has two different prime factorizations. For a further discussion of this point see ALGEBRAIC NUMBERS [IV.1 §§4–8].

What this example shows is that any proof of the fundamental theorem of arithmetic must use some feature of \mathbb{Z} , the set of integers, that is lacking in $\mathbb{Z}(\sqrt{-5})$. Since addition and multiplication work in a very similar way in both structures, it is not very easy to find such a feature, or at least not one that is relevant. It turns out that the important property that $\mathbb{Z}(\sqrt{-5})$ does not have is an appropriate analogue of the following basic principle for integers: that if m and n are integers, then one can write $n = qm + r$ with $0 \leq r < |m|$. This fact underlies EUCLID'S ALGORITHM [III.22], which plays an important role in the most commonly given proof of unique factorization.

V.17 The Fundamental Theorem of Calculus

The idea that integration is the “reverse” of differentiation is a familiar one. The fundamental theorem of calculus is the mathematical theorem that expresses this idea precisely. It has two parts: one says that differentiation “undoes” integration, and the other says that integration “undoes” differentiation. In order to formulate these principles properly, one must be careful to specify conditions on the function that is integrated or differentiated. The first part states the following: if f is

a continuous function defined on an interval $[a, b]$ and F is defined by the formula $F(x) = \int_a^x f(t) dt$, then F is differentiable and $F'(x) = f(x)$ for every x . As for the second, it states that if f is a differentiable function on $[a, b]$ and if its derivative f' is continuous, then $f(x) = f(a) + \int_a^x f'(t) dt$ for every x . It is important to realize that these statements are not true by definition. See [I.3 §5.5] for further discussion.

Gauss's Law of Quadratic Reciprocity

See FROM QUADRATIC RECIPROCITY TO
CLASS FIELD THEORY [V.31]

V.18 Gödel's Theorem

Peter J. Cameron

In response to problems in the foundations of mathematics such as *Russell's paradox* ("consider the set of all sets which are not members of themselves; is it a member of itself?"), HILBERT [VI.63] proposed that the consistency of any given part of mathematics should be established by finitary methods that could not lead to a contradiction. Any part for which this had been done could then be used as a secure foundation for all of mathematics.

An example of a "part of mathematics" is the arithmetic of the natural numbers, which can be described in terms of FIRST-ORDER LOGIC [IV.23 §1]. We begin with symbols, both logical (connectives such as "not" and "implies," quantifiers such as "for all," the equality symbol, symbols for variables, and punctuation) and non-logical (symbols for constants, relations, and functions suitable for the branch of mathematics under consideration). *Formulas* are finite strings of symbols built according to certain precise rules (which allow them to be mechanically recognized). We fix a certain set of formulas as our *axioms*, and we also choose a few *rules of inference* that allow us to infer some formulas from others. An example of a rule of inference is *modus ponens*: if we have inferred ϕ and $(\phi \rightarrow \psi)$, then we can infer ψ . A *theorem* is a formula that is at the end of a chain (or tree) of inferences that starts with axioms.

Axioms for the natural numbers were given by PEANO [VI.62] (see THE PEANO AXIOMS [III.69]). The nonlogical symbols are zero, the "successor function" s , addition, and multiplication. (The last two can be defined in terms of the others by inductive axioms: for example, the rules $x + 0 = x$ and $x + s(y) = s(x + y)$ define

addition.) The crucial axiom is the *principle of induction*, which asserts that if $P(n)$ is a formula such that $P(0)$ is true and $P(n)$ implies $P(s(n))$ for all n , then $P(n)$ is true for all n . Hilbert's specific challenge was to give a formal proof of the consistency of this theory: that is, a proof that no contradiction can be deduced from the axioms by the rules of first-order logic.

Hilbert's program was undone by two remarkable *incompleteness theorems* proved by GÖDEL [VI.92]. The first theorem states the following.

There are (first-order) statements about the natural numbers that can be neither proved nor disproved from Peano's axioms.

(This is sometimes qualified by being prefixed with, "If Peano's axioms are consistent, then..." However, since we accept the existence of the natural numbers, we do know that Peano's axioms are consistent, as the natural numbers model them. So the qualification is unnecessary here, although it would need to be included if we were discussing some axioms whose consistency was not clear.)

Gödel's proof is long, but it is based on two simple ideas. The first is *Gödel numbering*, which is a means of encoding each formula or sequence of formulas as a natural number in a systematic and mechanical way.

It can be shown that there is a two-variable formula $\pi(x, y)$ such that $\pi(m, n)$ holds if and only if " n is a proof of m ," which is a shorthand way of saying that m is the Gödel number of a formula ϕ and n is the Gödel number of a string of formulas that constitutes a proof of ϕ . Slightly more elaborately, there is a formula $\omega(x, y)$ such that $\omega(m, n)$ holds if and only if m is the Gödel number of a formula ϕ that has one free variable and n is the Gödel number of a proof of $\phi(m)$. (A free variable is one that is not quantified over. For example, $\phi(x)$ might be the formula $(\exists y)y^2 = x$, in which case x is the free variable. For this choice of ϕ , the number n would be the Gödel number of a proof that the Gödel number of ϕ was a perfect square.)

Now let $\psi(x)$ be the formula $(\forall y)(\neg \omega(x, y))$. If ϕ is a formula (with one free variable) with Gödel number m , then $\psi(m)$ tells us that there is no proof of $\phi(m)$. (It tells us this indirectly: what it actually says is that there is no y that is the Gödel number of such a proof.) Let p be the Gödel number of ψ itself, and let ζ be the formula $\psi(p)$.

This brings us to the second idea in the proof: *self-reference*. The formula ζ is carefully devised so that it asserts its own unprovability, since $\psi(p)$ tells us that

there is no proof of the formula with Gödel number $\phi(p)$, where ϕ is the formula with Gödel number p . In other words, it tells us that there is no proof of $\psi(p)$. Since ζ asserts its own unprovability, it must be unprovable (since a proof of ζ would be a proof that ζ had no proof, which is absurd). Since ζ asserts its unprovability and is unprovable, it is true, and since it is true it cannot be disprovable. (One might wonder why this argument that ζ is true does not constitute a proof of ζ . The answer is that although it is a rigorous demonstration of the truth of ζ , it is not a proof in Peano arithmetic. That is, it is not an argument that starts from the Peano axioms and uses the rules of inference of the kind we discussed earlier.)

Gödel numbering also allowed Gödel to consider the consistency of the axioms as a first-order formula: namely $(\forall y)(\neg(\pi(m, y)))$, where m is the Gödel number of the formula $0 = s(0)$ (or any other contradiction). Here is Gödel's second theorem.

It is impossible to prove from Peano's axioms that they are consistent.

The proofs of these theorems are not specific to the Peano axioms, but apply to any (consistent) system of mechanically recognizable axioms that is powerful enough to describe the natural numbers. Thus, completeness cannot be restored simply by adding a true but unprovable statement as a new axiom, for the resulting system is still strong enough for Gödel's theorem to apply to it.

It might seem that we could obtain a complete axiomatization of the natural numbers by simply taking all true statements as axioms. However, one requirement for Gödel's theorems is that the axioms should be recognizable by some mechanical method. (This is needed to construct the formula $\pi(x, y)$ at the start of the proof.) Indeed, we can deduce from this that (as subsequently pointed out by TURING [VI.94]) the true statements about the natural numbers *cannot* be mechanically recognized (that is, their Gödel numbers do not form a *recursive set*).

Gödel's true but unprovable statement is important for the foundations of mathematics, but it has no intrinsic interest in its own right. Later, Paris and Harrington gave the first example of a mathematically significant statement that is unprovable from Peano's axioms. Their statement is a variant of RAMSEY'S THEOREM [IV.19 §2.2]. Subsequently, many other "natural incompletenesses" have been found.

Of course, the consistency of Peano's axioms can be proved in a stronger system, since we could just add the (unprovable) consistency statement. Less trivially, since a model of the natural numbers can be constructed within set theory, the consistency of Peano arithmetic can be proved from THE ZERMELO-FRAENKEL AXIOMS [IV.22 §3.1] (known as ZFC) for set theory. Of course, ZFC cannot prove its own consistency, but the consistency of ZFC can be deduced from a yet stronger system (for example, adding an axiom that asserts the existence of a suitably "large" cardinal number such as an INACCESSIBLE CARDINAL [IV.22 §6]).

For small enough parts of mathematics, it is sometimes possible to find complete axiom systems (that is, systems that allow one to prove every true statement). For instance, this can be done for the theory of the natural numbers with zero, the successor function, and addition alone. Thus, multiplication is essential to Gödel's argument.

It is more elementary to see that Peano's axioms are not *categorical*: there are models for the axioms that are not isomorphic to the natural numbers. Such *non-standard models of arithmetic* contain infinitely large numbers (that is, numbers that are larger than all natural numbers).

Gödel's theorem has been a battleground for philosophers arguing about whether the human brain is a deterministic machine (in which case, presumably, we would not be able to prove any formally unprovable statement). Fortunately, there is not enough space in this article for more details!

The Goldbach Conjecture

See PROBLEMS AND RESULTS IN ADDITIVE
NUMBER THEORY [V.30]

V.19 Gromov's Polynomial-Growth Theorem

If G is a group and g_1, \dots, g_k are generators of G (meaning that every element of G can be expressed as a product of the g_i and their inverses), then we can define a *Cayley graph* by taking the elements of G as vertices and joining g to h if there is some i such that h is equal either to gg_i or to gg_i^{-1} .

For each r , let y_r be the number of elements that are at a distance of at most r from the identity: that is, the number of elements that can be written as a "word" of length at most r in the generators and their inverses.

(For instance, if $g = g_1 g_4 g_2^{-3}$, then we know that g belongs to γ_5 .) It turns out that if G is an infinite group, then the rate of growth of the sizes of the sets γ_r can tell one a great deal about G ; this is particularly true when the growth is less than exponential. (The growth is always bounded above by an exponential function, since there are at most exponentially many words of a given length in the generators g_1, \dots, g_r .)

If G is an Abelian group generated by g_1, \dots, g_k , then every element of γ_r is of the form $\sum_{i=1}^k a_i g_i$, where a_1, \dots, a_k are integers such that $\sum_{i=1}^k |a_i| \leq r$. It follows easily that the size of γ_r is at most $(2r+1)^k$ (and with a bit more effort one can improve this bound). Thus, as r tends to infinity, the growth rate of γ_r is bounded above by a polynomial of degree k in r . If G is the FREE GROUP [IV.10 §2] generated by g_1, \dots, g_k , then all words of length r in the elements g_i (but not their inverses) give rise to distinct elements of G , so the size of γ_r is at least k^r . Thus, in this case the growth rate is exponential. More generally, there will be an exponential growth rate whenever G contains a non-Abelian free subgroup.

These observations suggest that the growth rate is likely to be smaller if G is more like an Abelian group. Gromov's theorem is a remarkably precise result along these lines. It states that the growth rate of the sets γ_r is bounded above by a polynomial in r if and only if G has a nilpotent subgroup of finite index. This condition does indeed say that G is somewhat like an Abelian group, since nilpotent groups are "close to Abelian" and a subgroup of finite index is "close to the whole group." For example, a typical nilpotent group is the *Heisenberg group*, which consists of all 3×3 matrices with 0s below the diagonal, 1s on the diagonal, and integers above the diagonal. Given any two such matrices X and Y , the products XY and YX differ only in the top right-hand corner, and the "error matrix" $XY - YX$ commutes with everything in the group. In general, a nilpotent group is built out of Abelian groups in a controlled manner in a finite number of steps.

A fuller discussion of the theorem, including the exact definition of "nilpotent," can be found in GEOMETRIC AND COMBINATORIAL GROUP THEORY [IV.10]. Here we highlight the fact that it is a beautiful example of a *rigidity theorem*: if a group behaves roughly in the way that a nilpotent group would (because the growth rate of the sets γ_r is polynomial), then it must in fact be related to a nilpotent group in a very precise and algebraic way. (See MOSTOW'S STRONG RIGIDITY THEOREM [V.26] for another example of such a theorem.)

V.20 Hilbert's Nullstellensatz

Let f_1, \dots, f_n be a collection of polynomials in d complex variables z_1, \dots, z_d . Suppose that it is possible to find another collection of polynomials g_1, \dots, g_n such that

$$f_1(z)g_1(z) + f_2(z)g_2(z) + \dots + f_n(z)g_n(z) = 1$$

for every complex d -tuple $z = (z_1, \dots, z_d)$. Then it follows immediately that no such d -tuple can be a root of every single f_i , since otherwise the left-hand side would equal 0. Remarkably, the converse also holds: that is, if there is no d -tuple for which the polynomials f_i all vanish simultaneously, then it is possible to find polynomials g_i such that the above identity holds. This result is known as the *weak Nullstellensatz*.

A short (but clever) argument can be used to deduce *Hilbert's Nullstellensatz* from the weak Nullstellensatz. This again is a statement where a condition that is obviously necessary turns out to be sufficient. Suppose that h is another polynomial in d complex variables, that r is a positive integer, and that the polynomial h^r can be written in the form $f_1 g_1 + f_2 g_2 + \dots + f_n g_n$ for some collection of polynomials g_1, \dots, g_n . It follows immediately that $h(z) = 0$ whenever $f_i(z) = 0$ for every i . Hilbert's Nullstellensatz states that if $h(z) = 0$ whenever $f_i(z) = 0$ for every i , then there must be some positive integer r and some collection of polynomials g_1, \dots, g_n such that $h^r = f_1 g_1 + f_2 g_2 + \dots + f_n g_n$.

Hilbert's Nullstellensatz is discussed further in ALGEBRAIC GEOMETRY [IV.4 §§5, 12].

V.21 The Independence of the Continuum Hypothesis

The real numbers are UNCOUNTABLE [III.11], but do they form the "smallest" uncountable set? Equivalently, is it the case that if A is any set of real numbers, then either A is countable or there is a bijection between A and the set of all real numbers? The *continuum hypothesis* (or *CH*) is the assertion that this is indeed true. The notions of countability and uncountability were invented by CANTOR [VI.54], who was also the first to formulate CH. He tried hard to prove or disprove it, as did many others after him, but nobody succeeded.

Gradually, mathematicians came to entertain the idea that CH might be "independent" of normal mathematics: that is, independent of the usual ZFC AXIOMS

[IV.22 §3.1] of set theory. This would mean that it could be neither proved nor disproved from the ZFC axioms.

The first result in this direction was due to GÖDEL [VI.92], who showed that CH could not be *disproved* from the usual axioms. In other words, one could not reach a contradiction by assuming CH. To do this, he showed that inside every MODEL OF SET THEORY [IV.22 §3.2] there is a model in which CH holds. This model is called the “constructible universe.” Roughly speaking, it consists just of those sets that “have to exist” if the axioms are true. So, in this model, the set of reals is as small as it could possibly be. The “smallest uncountable size” is usually denoted \aleph_1 , and in Gödel’s construction the reals appear in \aleph_1 stages, with only countably many reals appearing at each stage. From this one can deduce that the number of reals is \aleph_1 , which is precisely the assertion of CH.

The other direction had to wait thirty years, until Paul Cohen invented the method of *forcing*. How would we make CH false? Starting from some model of set theory (in which CH might well hold), we would like to “add” some reals to it. Indeed, we would like to add enough that there are now more than \aleph_1 of them. But how do we “add” a real? We need to ensure that what we end up with is still a model of set theory, which is hard enough, but also that when we add new reals we do not alter the value of \aleph_1 (since otherwise the statement “the number of reals is \aleph_1 ” may still be true in the new model). This is an extremely complicated task, both conceptually and technically. See SET THEORY [IV.22] for more details about how it is carried out.

V.22 Inequalities

Let x and y be two nonnegative real numbers. Then $(\sqrt{x} - \sqrt{y})^2 = x + y - 2\sqrt{xy}$ is a nonnegative real number, from which it follows that $\frac{1}{2}(x + y) \geq \sqrt{xy}$. That is, the *arithmetic mean* of x and y is at least as big as the *geometric mean*. This conclusion is a very simple example of a mathematical inequality; its generalization to n numbers is called the *AM–GM inequality*.

In any branch of mathematics that has even the slightest flavor of analysis, inequalities will be of great importance: as well as analysis itself, this includes probability, and parts of combinatorics, number theory, and geometry. Inequalities are less prominent in some of the more abstract parts of analysis, but even there one needs them as soon as one wishes to apply the abstract results. For instance, one may not always need an inequality to prove a theorem

about continuous LINEAR OPERATORS [III.52] between BANACH SPACES [III.64], but the statement that some specific linear operator between two specific Banach spaces is continuous is an inequality, and often a very interesting one. We do not have space to discuss more than a small handful of inequalities in this article, but we shall include some of the most important ones in the toolbox of any analyst.

Jensen’s inequality is another fairly simple but useful inequality. A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is called *convex* if $f(\lambda x + \mu y) \leq \lambda f(x) + \mu f(y)$ whenever λ and μ are nonnegative real numbers with $\lambda + \mu = 1$. Geometrically, this says that all chords of the graph of the function lie above the graph. A straightforward inductive argument can be used to show that this property implies the same property for n numbers:

$$f(\lambda_1 x_1 + \cdots + \lambda_n x_n) \leq \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n)$$

whenever all the λ_i are nonnegative and $\lambda_1 + \cdots + \lambda_n = 1$. This is Jensen’s inequality.

The second derivative of the EXPONENTIAL FUNCTION [III.25] is positive, from which it follows that the exponential function itself is convex. If a_1, \dots, a_n are positive real numbers and we apply Jensen’s inequality to the numbers $x_i = \log(a_i)$, then we find, using standard properties of exponentials and LOGARITHMS [III.25 §4], that

$$a_1^{\lambda_1} \cdots a_n^{\lambda_n} \leq \lambda_1 a_1 + \cdots + \lambda_n a_n.$$

This is called the *weighted AM–GM inequality*. When all the λ_i are equal to $1/n$ it reduces to the usual AM–GM inequality. Applying Jensen’s inequality to other well-known convex functions produces several other well-known inequalities. For instance, if we apply it to the function x^2 , we obtain the inequality

$$(\lambda_1 x_1 + \cdots + \lambda_n x_n)^2 \leq \lambda_1 x_1^2 + \cdots + \lambda_n x_n^2, \quad (1)$$

which can be interpreted as saying that if X is a RANDOM VARIABLE [III.73 §4] on a finite sample space, then $(\mathbb{E}X)^2 \leq \mathbb{E}X^2$.

The *Cauchy–Schwarz inequality* is perhaps the most important inequality in all of mathematics. Suppose that V is a real vector space with an INNER PRODUCT [III.37] $\langle \cdot, \cdot \rangle$ on it. One of the properties of an inner product is that $\langle v, v \rangle \geq 0$ for every $v \in V$, with equality if and only if $v = 0$. Let us write $\|v\|$ for $\langle v, v \rangle^{1/2}$. If x and y are any two vectors in V with $\|x\| = \|y\| = 1$, then $0 \leq \|x - y\|^2 = \langle x - y, x - y \rangle = \langle x, x \rangle + \langle y, y \rangle - 2\langle x, y \rangle = 2 - 2\langle x, y \rangle$. It follows that $\langle x, y \rangle \leq 1 = \|x\| \|y\|$. Moreover, equality holds only if

$x = y$. We can obtain a general pair of vectors by multiplying x by λ and y by μ , for some nonnegative real numbers λ and μ . Then both sides of the inequality scale up by a factor of $\lambda\mu$, so we can conclude that the inequality $\langle x, y \rangle \leq \|x\| \|y\|$ holds in general, with equality if and only if x and y are proportional.

Particular inner-product spaces lead to special cases of this inequality, which are themselves often referred to as the Cauchy-Schwarz inequality. For instance, if we take the space \mathbb{R}^n with the inner product $\langle a, b \rangle = \sum_{i=1}^n a_i b_i$, then we obtain the inequality

$$\sum_{i=1}^n a_i b_i \leq \left(\sum_{i=1}^n a_i^2 \right)^{1/2} \left(\sum_{i=1}^n b_i^2 \right)^{1/2}. \quad (2)$$

It is not hard to deduce a similar inequality for complex scalars: one needs to replace a_i^2 and b_i^2 by $|a_i|^2$ and $|b_i|^2$ on the right-hand side. It is also not too hard to prove that inequality (2) is equivalent to the inequality (1) above.

Hölder's inequality is an important generalization of the Cauchy-Schwarz inequality. Again it has several versions, but the one that corresponds to inequality (2) is

$$\sum_{i=1}^n a_i b_i \leq \left(\sum_{i=1}^n |a_i|^p \right)^{1/p} \left(\sum_{i=1}^n |b_i|^q \right)^{1/q},$$

where p belongs to the interval $[1, \infty]$ and q is the *conjugate index* of p , which is defined to be the number that satisfies the equation $(1/p) + (1/q) = 1$. (We interpret $1/\infty$ to be 0.) If we write $\|a\|_p$ for the quantity $(\sum_{i=1}^n |a_i|^p)^{1/p}$, then this inequality can be rewritten in the succinct form $\langle a, b \rangle \leq \|a\|_p \|b\|_q$.

It is a straightforward exercise to find, for each sequence a , another (nonzero) sequence b such that equality occurs in the above inequality. Also, both sides of the inequality scale in the same way if you multiply b by a nonnegative scalar. It follows that $\|a\|_p$ is the maximum of $\langle a, b \rangle$ over all sequences b such that $\|b\|_q = 1$. Using this fact, it is easy to verify that the function $a \mapsto \|a\|_p$ satisfies *Minkowski's inequality*: $\|x + y\|_p \leq \|x\|_p + \|y\|_p$.

This gives some idea of why Hölder's inequality is so important. Once one has Minkowski's inequality, it is very easy to check that $\|\cdot\|_p$ is (as the notation suggests) a NORM [III.64] on \mathbb{R}^n . This is an even more basic example of the phenomenon mentioned at the beginning of the article: just to show that a certain normed space is a normed space, we have had to prove an inequality about real numbers. In particular, looking at the case $p = 2$, we see that the entire theory of HILBERT SPACES [III.37] depends on the Cauchy-Schwarz inequality.

Minkowski's inequality is a particular case of the *triangle inequality*, which states that if x , y , and z are three points in a METRIC SPACE [III.58], then $d(x, z) \leq d(x, y) + d(y, z)$, where $d(a, b)$ denotes the distance between a and b . When put like this, the triangle inequality is a tautology, since it is one of the axioms of a metric space. However, the statement that a particular notion of distance actually is a metric is far from vacuous. If our space is \mathbb{R}^n and we define $d(a, b)$ to be $\|a - b\|_p$, then Minkowski's inequality is easily seen to be equivalent to the triangle inequality for this notion of distance.

The inequalities above have natural “continuous analogues” as well. For example, here is a continuous version of Hölder's inequality. For two functions f and g defined on \mathbb{R} , let $\langle f, g \rangle$ be defined to be $\int_{-\infty}^{\infty} f(x)g(x) dx$, and write $\|f\|_p$ for the quantity $(\int_{-\infty}^{\infty} |f(x)|^p dx)^{1/p}$. Then, once again, $\langle f, g \rangle \leq \|f\|_p \|g\|_q$, where q is the conjugate index of p . Another example is a continuous version of Jensen's inequality, which states, in a continuous setting, that if f is convex and X is a random variable, then $f(\mathbb{E}X) \leq \mathbb{E}f(X)$.

In all the inequalities we have so far mentioned, we have been comparing two quantities A and B , and it has been easy to identify the extreme cases where the ratio of A to B is maximized. However, not all inequalities are like this. Consider, for instance, the following two quantities associated with a sequence of real numbers $a = (a_1, a_2, \dots, a_n)$. The first is the norm $\|a\|_2 = (\sum_{i=1}^n a_i^2)^{1/2}$. The second is the average of $|\sum_{i=1}^n \epsilon_i a_i|$ over all the 2^n sequences $(\epsilon_1, \epsilon_2, \dots, \epsilon_n)$ such that each ϵ_i is 1 or -1 . (In other words, for each i you randomly decide whether to multiply a_i by -1 or not, add up the results, and take the expected absolute value of the sum.) It is not the case that the first quantity is always less than the second. For instance, let $n = 2$, and let $a_1 = a_2 = 1$. Then the first quantity is $\sqrt{2}$ and the second is 1. However, *Khinchin's inequality* (or to be more accurate an important special case of Khinchin's inequality) is the remarkable statement that there is a constant C such that the first quantity is never more than C times the second. It is not hard to prove, using the inequality $\mathbb{E}X^2 \geq (\mathbb{E}X)^2$, that the first quantity is always at *least* as big as the second; so the two rather different looking quantities are in fact “equivalent, up to a constant.” But what is the best constant? In other words, how much bigger can the first quantity be than the second? This question was not answered until 1976, by Stanislaw Szarek, over fifty years after Khinchin proved the original inequality. The answer

turns out to be that the example given earlier is the extreme one: the ratio can never exceed $\sqrt{2}$.

This situation is typical. Another famous inequality for which the best constant was discovered much later than the inequality itself is the *Hausdorff-Young inequality*, which relates norms of functions with norms of their FOURIER TRANSFORMS [III.27]. Suppose that $1 \leq p \leq 2$, and that f is a function from \mathbb{R} to \mathbb{C} with the property that the norm

$$\|f\|_p = \left(\int_{-\infty}^{\infty} |f(x)|^p dx \right)^{1/p}$$

exists and is finite. Let \hat{f} be the Fourier transform of f and let q be the conjugate index of p . Then $\|\hat{f}\|_q \leq C_p \|f\|_p$ for some constant C_p that depends on p only (and not on f). Again, it was an open problem for many years to determine the best constant C_p . Some idea of why it might have been difficult can be gleaned from the fact that the “extreme” functions in this case are Gaussians: that is, functions of the form $f(x) = e^{-(x-\mu)^2/2\sigma^2}$. A sketch of a proof of the Hausdorff-Young inequality can be found in HARMONIC ANALYSIS [IV.11 §3].

There is an important class of inequalities known as *geometric inequalities*, where the quantities that are being compared are parameters associated with geometric objects. A famous example of such an inequality is the *Brunn-Minkowski inequality*, which states the following. Let A and B be two subsets of \mathbb{R}^n , and define $A + B$ to be the set $\{x + y : x \in A, y \in B\}$. Then

$$(\text{vol}(A + B))^{1/n} \geq \text{vol}(A)^{1/n} + \text{vol}(B)^{1/n}.$$

Here, $\text{vol}(X)$ denotes the n -dimensional volume (or, more formally, the LEBESGUE MEASURE [III.57]) of the set X . The Brunn-Minkowski inequality can be used to prove the equally famous *isoperimetric inequality* in \mathbb{R}^n (which is one of a large class of isoperimetric inequalities). Informally, this states that, of all sets with a given volume, the one with the smallest surface area is a sphere. An explanation of why this follows from the Brunn-Minkowski inequality can be found in HIGH-DIMENSIONAL GEOMETRY AND ITS PROBABILISTIC ANALOGUES [IV.26 §3].

We finish this brief sample with one further inequality, the *Sobolev inequality*, which is important in the theory of partial differential equations. Suppose that f is a differentiable function from \mathbb{R}^2 to \mathbb{R} . We can visualize its graph as a smooth surface in \mathbb{R}^3 lying above the x - y -plane. Suppose also that f is *compactly supported*, which means that there exists an M such that $f(x, y) = 0$ if the distance from (x, y) to $(0, 0)$ is

greater than M . We would now like to bound the size of f , as measured by some L_p norm, in terms of the size of its GRADIENT [I.3 §5.3] ∇f , as measured by some other L_p norm. The L_p norm of a function F is defined here as

$$\|f\|_p = \left(\int_{\mathbb{R}^2} |F(x, y)|^p dx dy \right)^{1/p}.$$

In one dimension, it is clear that no such bound is possible. For instance, we could have a differentiable function that was 1 everywhere on the interval $[-M, M]$, 0 everywhere outside the interval $[-(M+1), M+1]$, and gently decaying from 1 to 0 in between. Then if we widened the interval we would not change the size of the derivative: we would just move the two nonzero parts of the derivative further apart. On the other hand, using this widening we could increase the size of f as much as we liked. However, we cannot do this sort of construction in two dimensions, because now the “boundary” of the function increases as the size of the function increases. The Sobolev inequality tells us that if $1 \leq p < 2$ and $r = 2p/(2-p)$, then $\|f\|_r \leq C_p \|\nabla f\|_p$. To see why this might be reasonable, consider the case $p = 1$, so that $r = 2$. Let f be a function that is 1 everywhere inside the circle of radius M about the origin and 0 everywhere outside the circle of radius $M+1$. Then as M increases, the norm $\|f\|_2$ increases in proportion to M (since $\|f\|_2^2$ is approximately equal to the area of the circle of radius M), and so does $\|\nabla f\|_1$ (since it is roughly proportional to the length of the boundary of the circle). As this informal argument suggests, there are close connections between the Sobolev inequality and the isoperimetric inequality in the plane. And like the isoperimetric inequality, the Sobolev inequality has an n -dimensional version for each n : it is the same result, except that now the condition is that $1 \leq p < n$, and r is equal to $np/(n-p)$.

V.23 The Insolubility of the Halting Problem

What does it mean to understand a certain area of mathematics completely? One possible answer is that you understand it when you can solve its problems *mechanically*. Consider, for instance, the following question. Jim is half the age of his mother, and in twelve years' time he will be three-fifths of her age. How old is his mother now? For a child who is just old enough to understand the concept of “three-fifths,” this is likely to be an impossibly difficult problem. A bright and slightly

older child may be able to solve it after some hard thought, which will probably include a certain amount of trial and error. But for anybody who has learned how to translate such problems into equations and who knows how to solve two simultaneous linear equations, the problem is utterly routine: let x be Jim's age and y his mother's; then the problem tells us that $2x = y$ and $5(x + 12) = 3(y + 12)$; the second equation can be rearranged to give $3y - 5x = 24$; substituting $y = 2x$ gives $x = 24$, so $y = 48$.

The more mathematics one learns, the more one finds that problems that once seemed to be difficult and to require ingenuity have become routine in this sort of way, and it is eventually tempting to ask whether *all* of mathematics might, ultimately, be reducible to a mechanical procedure. And even if you think that that is a bit much to hope for, you can still ask the question about certain natural classes of problems, such as simultaneous linear equations. Perhaps there is always a mechanical procedure for solving the problems in any sufficiently "natural" class, even if there is not necessarily a systematic way of finding the mechanical procedure.

One class of problems that has been intensively studied for several centuries is that of *Diophantine equations*, which are equations in one or more variables where one stipulates that the solutions should be integers. The most famous Diophantine equation is the Fermat equation $x^n + y^n = z^n$ (see FERMAT'S LAST THEOREM [V.12]), but this is somewhat complicated because one of the variables, n , appears as an exponent. Suppose we restrict attention to *polynomial* equations, such as $x^2 - xy + y^2 = 157$. Is there a systematic way of telling whether such an equation has integer solutions?

The left-hand side of the equation $x^2 - xy + y^2 = 157$ is equal to $(x^2 + y^2 + (x - y)^2)/2$. Therefore, any solution (x, y) must satisfy $x^2 + y^2 \leq 314$, which makes it a short task to search through all possibilities until one discovers the solution $x = 12$ and $y = 13$ (or vice versa). However, an exhaustive search is not always possible: consider, for example, the equation $2x^2 - y^2 = 1$. This is a special case of the *Pell equation*, discussed in ALGEBRAIC NUMBERS [IV.1 §1]. The Pell equation can be solved systematically, with the help of CONTINUED FRACTIONS [III.22], and this leads to a systematic solution of all polynomial equations of degree up to 2 in two variables.

By the end of the nineteenth century, these and many other Diophantine equations had been completely solved, but there was no single overarching

method that dealt with all of them. This state of affairs prompted HILBERT [VI.63] to include, as the tenth in his famous list of twenty-three unsolved problems, the question of whether there was a single, universal procedure for solving all polynomial Diophantine equations in any number of variables. Later, in 1928, he asked the more general question alluded to earlier: is there a universal procedure for determining the truth or falsity of any mathematical statement? This question became known as the *Entscheidungsproblem* (which means "decision problem" in German).

Hilbert expected, or at least hoped, that the answers to both questions would be yes. In other words, he hoped that the mathematicians of his day were in the position of the child who has not yet learned how to solve simultaneous equations. Perhaps a new age was dawning in which it would be possible, at least in principle, to solve all mathematical problems systematically and without relying on native wit.

The evidence in favor of such a view was not very strong: although problems of some kinds could be solved fully systematically, others, including Diophantine equations, stubbornly resisted, and the role of ingenuity in mathematical research appeared to be as important as ever. But if one wanted to give a *negative* answer to Hilbert's questions, then one faced a major challenge: in order to prove rigorously that there is no systematic procedure for accomplishing a particular task, one has to be absolutely clear about what a "systematic procedure" actually is.

Nowadays there is an easy answer to this: a systematic procedure is anything that you can program a computer to do. (Strictly speaking, this is an oversimplification, because one also makes the idealizing assumption that the computer has unlimited storage space.) Our feeling that we do not have to think too hard to solve simultaneous equations is reflected in the fact that we can devise a computer program to do it for us (though if we want the program to be fast and numerically robust, we will face very interesting problems: see NUMERICAL ANALYSIS [IV.21 §4]). However, Hilbert asked the questions before computers existed, so it was a remarkable achievement when in 1936 CHURCH [VI.89] and TURING [VI.94] independently managed to formalize the notion of what we now call an ALGORITHM [IV.20 §1]. That is, they each gave a precise definition of the notion of an algorithm. Their definitions were quite different, but later shown to be equivalent, which means that anything that can be done by an algorithm in Church's sense can be done by an algorithm

PUP: Tim prefers this sentence and the next as they are. OK?

in Turing's sense, and vice versa. Turing's formalization, which had a big influence on the design of modern computers, is discussed in COMPUTATIONAL COMPLEXITY [IV.20 §1.1], while Church's is described in ALGORITHMS [II.4 §3.2], but for the purposes of this article we shall use the anachronistic definition with which this paragraph began.

It turns out that once one has *any* sufficiently precise notion of "algorithm," one is just a few short steps away from a negative answer to Hilbert's Entscheidungsproblem. To see this, imagine that L is some programming language (such as Pascal or C++). Given any string of symbols, we can ask of it the following question: if I present that string of symbols to my computer as a program in L , will the program run forever, or will it eventually stop? This is called the *halting problem*. (Note that the word "problem" really means "class of problems.") The halting problem may not seem very mathematical, but certain instances of it certainly are. For example, suppose that after a quick look at a program you realize that it does the following. In one portion of the memory it stores an even number n , which at the beginning is set to 6. It then checks for every odd number m less than n whether m and $n - m$ are both prime. If the answer is yes for some m , then it adds 2 to n and repeats. If the answer is no for all m , then it halts. This program will halt if and only if the Goldbach conjecture (see PROBLEMS AND RESULTS IN ADDITIVE NUMBER THEORY [V.30]) is false.

Turing proved that *there is no systematic procedure for solving the halting problem*. (Church proved an analogous result for his notion of *recursive functions*.) Let us see how Turing's argument works for the language L . In this case, it shows that there is no systematic procedure for recognizing which strings of symbols form programs in L that halt, and which do not. The proof is a *reductio ad absurdum*, so we begin by assuming that there *is* such a procedure. Let us call it P . Suppose that L is like most computer languages, in that a typical program asks for an *input*, which affects its subsequent behavior. Then P will be able to tell, given any pair of strings (S, I) , whether S is a program in L that halts if the input is I .

Now let us create a new procedure Q out of P . Given any string S , we start by getting Q to run P on the pair (S, S) . If P judges that S does *not* halt when presented with itself as input, we then cause Q to halt. But if P judges that S *does* halt when presented with itself as input, then we artificially send Q into an endless loop, so that it does not halt. (If S is not a valid program in L ,

then let us say that Q halts—it does not really matter though.) To summarize, if S halts for input S , then Q does not halt for S , and if S does not halt for S , then Q does halt for S .

But now let us suppose that S is the program for Q itself. Does Q halt with input S ? If it does, then S halts with input S , so Q does not halt. If it does not, then S does not halt with input S , so Q does halt. This is a contradiction, and therefore the procedure P out of which Q was built could not have existed.

That solves the general version of Hilbert's problem: there is no algorithm that will determine the truth or falsity of arbitrary mathematical statements. But it does so by constructing, for any given algorithm, a rather artificial statement. We do not yet have an answer to the question of what happens if we look at more specific and more natural classes of statements, such as that a given Diophantine equation has a solution.

Remarkably, however, specific questions of this kind can often be shown to be *equivalent* to the general question, by a technique known as *encoding*. For example, there is no algorithm that will take as its input a set of polygonal tiles (suitably represented) and tell you whether it is possible to tile the plane using copies of just those tiles. How do we know this? Well, given any algorithm, there is a clever way of devising a set of tiles (this is the encoding) that will tile the plane if and only if the algorithm halts. Therefore, if there were an algorithm for determining whether the tiles could tile the plane, then there would be an algorithm for solving the halting problem—but there is not.

Another famous example of a more specific problem for which there is no algorithm is the *word problem for groups*. Here you are given a set of generators and relations for a group and asked whether the group is trivial—that is, whether it contains just the identity. Again, an algorithm that could decide this would give us an algorithm that could solve the halting problem, so there cannot be one. The encoding process used to prove this is much more difficult than it is for tiling the plane: the insolubility of the word problem for groups is a famous theorem proved by Pyotr Novikov in 1952. For a much fuller explanation of this problem and its solution, see GEOMETRIC AND COMBINATORIAL GROUP THEORY [IV.10].

Finally, what about Hilbert's tenth problem? This has become another famous and very hard theorem, due to Yuri Matiyasevitch in 1970, who built on work of Martin Davis, Hilary Putnam, and Julia Robinson. Matiya-

sevitich managed to produce a system of ten equations, involving two parameters m and n , that could be solved in integers if and only if m was the $2n$ th Fibonacci number. From Robinson's work it followed that, given any algorithm with integer inputs, there was a system of Diophantine equations, involving a parameter q , that could be solved if and only if the algorithm failed to halt at q . That is, any instance of the halting problem can be encoded as a system of Diophantine equations, so there is no general algorithm for deciding whether Diophantine equations can be solved.

Different people draw different morals from these results. In the opinion of some mathematicians, they show that there will always be a place for human creativity in mathematics, however powerful the computers of the future might be. Others maintain that although we now know that we cannot systematically solve all problems in mathematics, the effect on most mathematics is very slight: one should be aware that certain kinds of problems are sometimes equivalent to the halting problem, and that is it. Still others point out that it is often easy to devise an algorithm to solve a problem but much harder to make it *efficient*. This issue is discussed in great detail in COMPUTATIONAL COMPLEXITY [IV.20].

Turing's argument for the insolubility of the halting problem is closely related to GÖDEL'S THEOREM [V.18], and both proofs use *diagonal arguments*, which are discussed in COUNTABLE AND UNCOUNTABLE SETS [III.11].

V.24 The Insolubility of the Quintic

Martin W. Liebeck

Every student will be familiar with the formula for the roots of a quadratic polynomial $ax^2 + bx + c$, namely $(-b \pm \sqrt{b^2 - 4ac})/2a$. Perhaps less familiar is the fact that there is also a formula for the roots of a cubic: write the cubic as $x^3 + ax^2 + bx + c$, and make the substitution $y = x + \frac{1}{3}a$ to rewrite it in the form $y^3 + hy + k$. The roots of this are then of the form

$$\sqrt[3]{\frac{1}{2}(-k + \sqrt{k^2 + 4h^3})} + \sqrt[3]{\frac{1}{2}(-k - \sqrt{k^2 + 4h^3})}.$$

While the quadratic formula was known to the Greeks, the cubic formula was not found until the sixteenth century. In the same century a formula for the roots of quartic (degree 4) polynomials was also found. The formulas for quadratics, cubics, and quartics all arise by applying a sequence of arithmetic operations (addition, subtraction, multiplication, division)

together with extraction of roots (square roots, cube roots, and so on) to the coefficients of the original polynomial. Such a formula is called a *radical* expression for the roots.

The next step, naturally enough, was the quintic (i.e., polynomial of degree 5). However, several hundred years passed without anyone finding a radical formula for the roots of a general quintic polynomial.

There was a good reason for this. There is no such formula. Nor is there a formula for polynomials of degree greater than 5. This fact was first established in the early nineteenth century by ABEL [VI.33] (who died aged twenty-six), after which GALOIS [VI.41] (who died aged twenty-one) built an entirely new theory of equations that not only explained the nonexistence of formulas but laid the foundations for a whole edifice of algebra and number theory known as *Galois theory*, a major area of modern-day research.

One of the key ideas of Galois was to associate with any polynomial $f = f(x)$ a GROUP [I.3 §2.1] $\text{Gal}(f)$ (the Galois group of f), which is a finite group that permutes the roots of f . This group is defined in terms of certain FIELDS [I.3 §2.2], which for these purposes can be thought of as subsets F of the COMPLEX NUMBERS [I.3 §1.5] \mathbb{C} having the property that if a, b are any two elements of F , then all the numbers $a + b$, $a - b$, ab , and a/b also lie in F (where we assume that $b \neq 0$ in the last case to avoid dividing by 0). The standard mathematical language for this property is to say that F is “closed under” the usual arithmetic operations of addition, subtraction, multiplication, and division. For example, the rationals \mathbb{Q} form a field, as does $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$ (this is clearly closed under addition, subtraction, and multiplication, and is also closed under division since $1/(a + b\sqrt{2}) = a/(a^2 - 2b^2) - b\sqrt{2}/(a^2 - 2b^2)$). A polynomial $f(x)$ of degree n with rational coefficients has n complex roots by THE FUNDAMENTAL THEOREM OF ALGEBRA [V.15]—call them $\alpha_1, \dots, \alpha_n$. The *splitting field* of f is defined to be the smallest field containing \mathbb{Q} and all the α_i , and is written as $\mathbb{Q}(\alpha_1, \dots, \alpha_n)$. For example, the polynomial $x^2 - 2$ has roots $\pm\sqrt{2}$, so its splitting field is $\mathbb{Q}(\sqrt{2})$, defined above. Less trivially, $x^3 - 2$ has roots $\alpha, \alpha\omega, \alpha\omega^2$, where $\alpha = 2^{1/3}$, the real cube root of 2, and $\omega = e^{2\pi i/3}$, so its splitting field is $\mathbb{Q}(\alpha, \omega)$, which consists of all complex numbers $a_1 + a_2\alpha + a_3\alpha^2 + a_4\omega + a_5\alpha\omega + a_6\alpha^2\omega$ with $a_i \in \mathbb{Q}$. (Notice that we do not have to include ω^2 in such expressions since $\omega^3 = 1$, so $(\omega - 1)(\omega^2 + \omega + 1) = \omega^3 - 1 = 0$, which implies that $\omega^2 = -\omega - 1$.)

Let $E = \mathbb{Q}(\alpha_1, \dots, \alpha_n)$ be the splitting field of our polynomial f . An *automorphism* of E is a bijection $\phi : E \rightarrow E$ that preserves addition and multiplication—in other words, $\phi(a + b) = \phi(a) + \phi(b)$ and $\phi(ab) = \phi(a)\phi(b)$ for all $a, b \in E$. Such a function necessarily also preserves subtraction and division, and fixes every rational number. Denote by $\text{Aut}(E)$ the set of all automorphisms of E . For example, when $E = \mathbb{Q}(\sqrt{2})$, any automorphism ϕ satisfies

$$2 = \phi(2) = \phi(\sqrt{2}\sqrt{2}) = \phi(\sqrt{2})\phi(\sqrt{2}) = \phi(\sqrt{2})^2,$$

and therefore $\phi(\sqrt{2}) = \sqrt{2}$ or $-\sqrt{2}$. In the first case $\phi(a + b\sqrt{2}) = a + b\sqrt{2}$ for all $a, b \in \mathbb{Q}$, while in the second $\phi(a + b\sqrt{2}) = a - b\sqrt{2}$. Both of these are automorphisms of E ; call them ϕ_1, ϕ_2 , so that $\text{Aut}(E) = \{\phi_1, \phi_2\}$.

The composition $\phi \circ \psi$ of two automorphisms ϕ, ψ of E is again an automorphism, and so is the inverse function ϕ^{-1} , while the identity function ι defined by $\iota(e) = e$ for all $e \in E$ is also an automorphism. Since composition of functions is an associative operation, it follows that $\text{Aut}(E)$ is a group under composition. Define the *Galois group* $\text{Gal}(f)$ of our polynomial $f(x)$ with splitting field E to be this group $\text{Aut}(E)$. Thus, for example, $\text{Gal}(x^2 - 2) = \{\phi_1, \phi_2\}$. Notice that ϕ_1 is the identity ι , while $\phi_2^2 = \phi_2 \circ \phi_2 = \phi_1$, so this is just a cyclic group of order 2. Similarly, if $f(x) = x^3 - 2$, with splitting field $E = \mathbb{Q}(\alpha, \omega)$ as above, then any $\phi \in \text{Aut}(E)$ satisfies $\phi(\alpha)^3 = \phi(\alpha^3) = \phi(2) = 2$, and therefore $\phi(\alpha) = \alpha, \alpha\omega$, or $\alpha\omega^2$; likewise $\phi(\omega) = \omega$ or ω^2 . Once $\phi(\alpha)$ and $\phi(\omega)$ are specified, ϕ is completely determined (since $\phi(a_1 + a_2\alpha + \dots + a_6\alpha^2\omega) = a_1 + a_2\phi(\alpha) + \dots + a_6\phi(\alpha)^2\phi(\omega)$), so there are just six possibilities for the automorphism ϕ . It turns out that each of these is indeed an automorphism, and therefore $\text{Gal}(x^3 - 2)$ is a group of order 6. In fact, this group is isomorphic to the SYMMETRIC GROUP [III.70] S_3 , as can be seen by considering each automorphism as a permutation of the three roots of $f(x)$.

Now that the Galois group is defined, it is possible to state some of Galois's fundamental results that lead to the insolubility of the quintic. Each subgroup H of $G = \text{Gal}(f)$ has a *fixed field* H^\dagger , which is defined to be the set of all numbers $a \in E$ such that $\phi(a) = a$ for all $\phi \in H$. Galois proved that the association between H and H^\dagger gives a one-to-one correspondence between subgroups of G and fields which lie between \mathbb{Q} and E (the so-called *intermediate subfields* of E). The condition that $f(x)$ has a radical formula for its roots leads to certain special kinds of intermediate subfields,

and hence to certain special subgroups of G , and eventually to Galois's most famous theorem: the polynomial $f(x)$ has a radical formula for its roots if and only if its Galois group $\text{Gal}(f)$ is a *soluble* group. (This means that $G = \text{Gal}(f)$ has a sequence of subgroups $1 = G_0 < G_1 < \dots < G_r = G$ such that for each i , G_i is a NORMAL SUBGROUP [I.3 §3.3] of G_{i+1} and the factor group G_{i+1}/G_i is Abelian.)

It follows from Galois's theorem that to demonstrate the insolubility of the quintic, it is enough to produce a quintic $f(x)$ such that $\text{Gal}(f)$ is not a soluble group. An example of such a quintic is $f(x) = 2x^5 - 5x^4 + 5$: one can show first that $\text{Gal}(f)$ is isomorphic to the symmetric group S_5 ; and second that S_5 is not a soluble group. Here is a brief sketch of how the argument goes. First one establishes that $f(x)$ is an irreducible polynomial (i.e., is not the product of two rational polynomials of smaller degree) with five distinct complex roots. Thus, as observed above, $\text{Gal}(f)$ can be regarded as a subgroup of S_5 that permutes the five roots. By sketching the graph of $f(x)$ one can easily see that three of its roots are real and that the other two, call them α_1 and α_2 , are complex conjugates of each other. Since the complex conjugation map $z \rightarrow \bar{z}$ always gives an automorphism in $\text{Gal}(f)$, it follows that $\text{Gal}(f)$ is a subgroup of S_5 that contains a 2-cycle, namely $(\alpha_1\alpha_2)$. Another basic general fact is that the Galois group of an irreducible polynomial permutes the roots *transitively*, meaning that for any two roots α_i, α_j there exists an automorphism in $\text{Gal}(f)$ that sends α_i to α_j . Thus, our group $\text{Gal}(f)$ is a subgroup of S_5 that permutes the five roots transitively and contains a 2-cycle. At this point some fairly elementary group theory shows that $\text{Gal}(f)$ must actually be the whole of S_5 . Finally, the fact that S_5 is not a soluble group follows easily from the fact that the alternating group A_5 is a non-Abelian simple group (i.e., it has no normal subgroups apart from the identity subgroup and A_5 itself).

These ideas can be extended to produce polynomials of any degree $n \geq 5$ that have Galois group S_n , and that are therefore not soluble by radicals. The reason this cannot be done for quartics, cubics, and quadratics is that S_4 and all its subgroups are soluble groups.

V.25 Liouville's Theorem and Roth's Theorem

One of the most famous theorems in mathematics is the statement that $\sqrt{2}$ is irrational. This means that there is no pair of integers p and q such that $\sqrt{2} = p/q$, or

equivalently that the equation $p^2 = 2q^2$ has no integer solutions apart from the trivial solution $p = q = 0$. The argument that proves this can be considerably generalized, and, in fact, if $P(x)$ is any polynomial with integer coefficients and leading coefficient 1, then all its roots are either integers or irrational numbers. For example, since $x^3 + x - 1$ is negative when $x = 0$ and positive when $x = 1$ it must have a root strictly between 0 and 1. This root is not an integer, so it must be irrational.

Once one has proved that a number is irrational, it may seem as though not much more can be said. However, this is very far from true: given an irrational number, one can ask *how close* it is to being rational, and fascinating and extremely difficult questions arise as soon as one does so.

It is not immediately obvious what this question means, since every irrational number can be approximated as closely as you like by rational numbers. For example, the decimal expansion of $\sqrt{2}$ begins 1.414213..., which tells us that $\sqrt{2}$ is within $1/100\,000$ of the rational number $141\,421/100\,000$. More generally, for any positive integer q we can let p be the largest integer such that $p/q < \sqrt{2}$, and then p/q will be within $1/q$ of $\sqrt{2}$. In other words, if we want an approximation to $\sqrt{2}$ with accuracy $1/q$, we can obtain it if we use a denominator of q .

However, we can now ask the following question: are there denominators q for which one can obtain an accuracy much better than $1/q$? The answer turns out to be yes. To see this, let N be a positive integer and consider the numbers $0, \sqrt{2}, 2\sqrt{2}, \dots, N\sqrt{2}$. Each of these can be written in the form $m + \alpha$, where m is an integer and α , the fractional part, lies between 0 and 1. Since there are $N + 1$ numbers, at least two of their fractional parts must be within $1/N$ of each other. That is, we can find integers $r < s$ between 0 and N such that if we write $r\sqrt{2} = n + \alpha$ and $s\sqrt{2} = m + \beta$, then $|\alpha - \beta| \leq 1/N$. Thus, if we set $\gamma = \alpha - \beta$, we have $(s - r)\sqrt{2} = n - m + \gamma$ and $|\gamma| \leq 1/N$. If we now let $q = s - r$ and $p = n - m$, then $\sqrt{2} = p/q + \gamma/q$, so $|\sqrt{2} - p/q| \leq 1/qN$. Since $N \geq q$, $1/qN \leq 1/q^2$, so for at least some positive integers q we can achieve an accuracy of $1/q^2$ using a denominator of q .

A different argument shows that we cannot do substantially better than this. Let p and q be any two positive integers. Since $\sqrt{2}$ is irrational, p^2 and $2q^2$ are distinct positive integers, which implies that $|p^2 - 2q^2| \geq 1$. On factorizing, we deduce that $|p - q\sqrt{2}|(p + q\sqrt{2}) \geq 1$. We can now divide through by q^2 and obtain the inequality $|p/q - \sqrt{2}|(p/q + \sqrt{2}) \geq 1/q^2$. We may

as well assume that p/q is less than 2, since otherwise it is not a good approximation to $\sqrt{2}$. But then $p/q + \sqrt{2}$ is less than 4, so the inequality implies that $|p/q - \sqrt{2}| \geq 1/4q^2$. Thus, with a denominator of q we cannot achieve an accuracy better than $1/4q^2$.

A generalization of this argument proves *Liouville's theorem*: if x is an irrational root of a polynomial of degree d and p and q are integers, then $|p/q - x|$ cannot be substantially smaller than $1/q^d$. When $x = \sqrt{2}$ this reduces to what we have just shown, since then $x^2 - 2 = 0$ and we can set $d = 2$. However, from Liouville's theorem we know many similar facts, such as that $|p/q - \sqrt[3]{2}|$ cannot be substantially smaller than $1/q^3$.

Roth's theorem, proved in 1955, is the astonishing assertion that the power d that appears in Liouville's theorem can be improved—almost as far as 2. To be precise, given any irrational root x of any polynomial, and any number $r > 2$, there is a constant $c > 0$ with the property that $|p/q - x|$ is always at least as big as c/q^r . (The proof gives no information whatsoever about c beyond the fact that it is positive. It is a major open problem to understand something about how c depends on r and x .)

To see why this is a much deeper result than Liouville's theorem, consider the example of $\sqrt[3]{2}$. Underlying the proof that $|p/q - \sqrt[3]{2}|$ is never much smaller than $1/q^3$ is the simple fact that p^3 and $2q^3$ are distinct integers and therefore differ by at least 1. In order to prove a substantially better result such as Roth's theorem, one must show much more: that p^3 and $2q^3$ differ by an amount that grows as p and q grow. For example, if one wishes to prove Roth's theorem when $r = \frac{5}{2}$, it is necessary to show that p^3 and $2q^3$ must always differ by an amount comparable to or greater than \sqrt{p} , and it is far from obvious why this should be so.

The Mordell Conjecture

See RATIONAL POINTS ON CURVES AND
THE MORDELL CONJECTURE [V.32]

V.26 Mostow's Strong Rigidity Theorem

David Fisher

1 What Are Rigidity Theorems?

A typical *rigidity theorem* is a statement that some class of objects is much smaller than one might expect. To

make this notion clear, let us look at some examples of MODULI SPACES [IV.8] that might lead us to expect that spaces of a certain type would in general be large.

2 Some Moduli Spaces

A METRIC [III.58] on an n -dimensional MANIFOLD [I.3 §6.9] is called *flat* if it is locally isometric to the usual metric on the Euclidean space \mathbb{R}^n . In other words, every point x in the manifold is contained in a neighborhood N_x such that there is a distance-preserving bijection from N_x to a subset of \mathbb{R}^n . For our first example, we shall consider flat metrics on a torus. We shall consider just the two-dimensional torus, but the phenomena we shall discuss occur in higher dimensions as well.

The simplest way of putting a flat metric on the two-dimensional torus \mathbb{T}^2 is to view it as the QUOTIENT [I.3 §3.3] of \mathbb{R}^2 by a discrete subgroup, or *lattice*, that is isomorphic to \mathbb{Z}^2 . In fact, it is not too hard to see that every flat metric arises in essentially this way. However, there is a choice involved: the choice of which lattice to take. An obvious choice is \mathbb{Z}^2 itself. But one can also take any invertible linear transformation A , apply it to \mathbb{Z}^2 , and then define the torus as $\mathbb{R}^2/A(\mathbb{Z}^2)$, which gives rise to another metric. A natural question to ask is, when do two choices of A give rise to the same metric? Usually, one studies only the cases when the DETERMINANT [III.15] of A is 1, since it is easy to deduce from these what happens in general. The group of all such linear maps is called $\mathrm{SL}_2(\mathbb{R})$.

If A is orthogonal, then it just rotates the lattice \mathbb{Z}^2 and therefore $A(\mathbb{Z}^2)$ gives rise to the same metric as \mathbb{Z}^2 . What is slightly less obvious is that there are other maps A that give rise to this metric as well, namely all maps of determinant 1 whose matrices with respect to the standard basis of \mathbb{R}^2 have integer entries. The group of all these maps is called $\mathrm{SL}_2(\mathbb{Z})$. If A belongs to $\mathrm{SL}_2(\mathbb{Z})$, then the reason that $A(\mathbb{Z}^2)$ gives rise to the same metric as \mathbb{Z}^2 is simple: $A(\mathbb{Z}^2)$ is actually equal to \mathbb{Z}^2 .

Loosely speaking, what we have just done is identify the space of flat metrics on \mathbb{T}^2 with the set $\mathrm{SL}_2(\mathbb{Z}) \backslash \mathrm{SL}_2(\mathbb{R}) / \mathrm{SO}(2)$. (This is notation for the set $\mathrm{SL}_2(\mathbb{R})$, with two maps A and B considered equivalent if B can be expressed as A multiplied by a product of matrices from $\mathrm{SO}(2)$ and $\mathrm{SL}_2(\mathbb{Z})$.) In higher dimensions, a similar discussion shows that one can identify the space of flat metrics on the n -dimensional torus \mathbb{T}^n with $\mathrm{SL}_n(\mathbb{Z}) \backslash \mathrm{SL}_n(\mathbb{R}) / \mathrm{SO}(n)$.

Returning to two dimensions, a torus is a surface of genus 1 (since it has one “hole”). A similar construction

gives rise to a moduli space of metrics on a surface of higher genus, but now the metrics will be *hyperbolic* rather than flat. The UNIFORMIZATION THEOREM [V.37] says that any compact connected surface admits a metric of constant CURVATURE [III.13]: when the genus is 2 or more, this curvature must be negative, which implies that the surface is a QUOTIENT [I.3 §3.3] of the HYPERBOLIC PLANE [I.3 §6.6] \mathbb{H}^2 by a group Γ that acts on \mathbb{H}^2 as a set of isometries. (See FUCHSIAN GROUPS [III.28].)

Conversely, if we want to construct a metric of constant curvature on a surface of higher genus, we can take a subgroup Γ of the group of isometries of \mathbb{H}^2 (which is isomorphic to $\mathrm{SL}_2(\mathbb{R})$) and we can consider the quotient \mathbb{H}^2/Γ , which is analogous to the quotient $\mathbb{R}^2/\mathbb{Z}^2$ that we considered earlier. If Γ has no elements of finite order and if for each x the *orbit* of x (the set of images of x under the isometries in Γ) is a discrete subset of \mathbb{H}^2 , then this space is a manifold. Furthermore, if there is a compact region in \mathbb{H}^2 , called a *fundamental domain*, whose translates cover \mathbb{H}^2 , then the manifold is compact. There are two fairly simple ways to construct examples of groups Γ with these properties: one is to use reflection groups and the other is to use a bit of number theory.

Now we can ask the same question for these metrics. In other words, given a surface S of genus at least 2, how many hyperbolic metrics can we find on S ? The answer is quite similar to the answer for \mathbb{T}^2 . For instance, if the genus is 2, then there is a connected six-dimensional space of such structures. This is a bit more difficult to see, as the space is not constructed in any simple way from a LIE GROUP [III.50 §1] (such as $\mathrm{SL}_n(\mathbb{R})$) and its subgroups. We will not describe this construction here but it can be found in Thurston (1997) or in MODULI SPACES [IV.8].

T&T note: Fisher would like to see a final proof when article finalised.

3 Mostow's Theorem

Thinking about the last two sets of examples leads to a natural question: what about compact three-dimensional hyperbolic manifolds? Or n -dimensional ones? To be clear, a compact n -dimensional hyperbolic manifold is the quotient of \mathbb{H}^n by a discrete group Γ of isometries of the hyperbolic n -space \mathbb{H}^n such that Γ has no elements of finite order and there is a compact fundamental domain for Γ . Given this description, the reader may wonder if there are any such groups Γ . Once again, there are two easy ways of constructing them, one using a bit of number theory and another using reflection groups. (However, slightly surprisingly, the method using reflection groups works

only in fairly small dimensions.) The constructions are all a bit technical so we will not go through them here. There are also many other examples of compact hyperbolic manifolds, particularly in three dimensions, where “most” manifolds are hyperbolic by the GEOMETRIZATION THEOREM [IV.7 §2.4].

Here we shall concentrate less on the existence of hyperbolic manifolds and more on the question that has been our principal concern in this article: if X is a manifold that can be represented in the form \mathbb{H}^n/Γ , then how many ways are there of giving X this structure? This question is equivalent to asking how many injective homomorphisms there are from Γ to the group of all isometries of \mathbb{H}^n such that the image of Γ is discrete and cocompact. (A subset X of a group G is *cocompact* if there is a compact subset K of G such that $XK = G$. For instance, \mathbb{Z}^2 is a cocompact subset of \mathbb{R}^2 because $\mathbb{R}^2 = \mathbb{Z}^2 + [0, 1]^2$ and the closed unit square $[0, 1]^2$ is compact.) As we have seen, when $n = 2$ there is a continuum of such homomorphisms, and the same is true in all dimensions if we replace \mathbb{H}^n by \mathbb{R}^n . So it is rather surprising that when $n \geq 3$, the answer for \mathbb{H}^n is exactly 1. This is a special case of Mostow’s rigidity theorem.

What does this result mean? Suppose we know that a manifold M is a quotient of \mathbb{H}^n by some discrete cocompact group of isometries. The topology of M completely determines the group Γ up to isomorphism: it is just the FUNDAMENTAL GROUP [IV.6 §2] of M . The result we have just stated tells us that this purely topological information about the manifold M completely determines the geometry of \mathbb{H}^n/Γ (that is, its structure as a metric space). More precisely, it says that any homeomorphism, or even homotopy equivalence, from M to another hyperbolic manifold N is homotopic to an isometry. In other words, any purely topological equivalence can be realized as a geometric equivalence.

The full Mostow rigidity theorem concerns objects called compact locally symmetric manifolds. Given a manifold with a metric, we say that it is *locally symmetric* if the *central symmetry* at every point is a local isometry. The central symmetry at a point m is defined formally as multiplication by -1 in the tangent space to m : one can picture it as taking a very small neighborhood of m and “reflecting through m .” It turns out that every locally symmetric space is a quotient of a *symmetric space*: that is, a space such that the central symmetry at every point is a global isometry. Clearly, symmetric spaces have very large isometry groups. The

work of CARTAN [VI.69] shows that the resulting isometry groups are exactly the semisimple LIE GROUPS [III.50 §1]. We will not say precisely what these are, but they include the classical matrix groups such as $SL_n(\mathbb{R})$, $SL_n(\mathbb{C})$, and $Sp_n(\mathbb{R})$. Other examples, which can also be realized as matrix groups, include the isometry groups of complex and quaternionic hyperbolic spaces.

In general, given a Lie group G and a discrete subgroup Γ , we say that Γ is a cocompact lattice if there is a compact fundamental domain for Γ in G . Cartan’s theorem has the consequence that any compact locally symmetric space is a quotient $\Gamma \backslash G/K$, where G is the isometry group of the universal cover and K is the (necessarily compact) set of isometries that fix a specified point. Mostow’s theorem says the same here as it said for \mathbb{H}^n/Γ : given such a manifold, there is only one way to realize it as $\Gamma \backslash G/K$. Or, equivalently, any homeomorphism between two such manifolds is always homotopic to an isometry unless the relevant locally symmetric space is a product of a flat torus or a hyperbolic surface with some other locally symmetric manifold.

One might well ask how Mostow discovered such a phenomenon. His work certainly did not occur in a vacuum. In fact, earlier work of Calabi, Selberg, Vesentini, and WEIL [VI.93] had already shown that the moduli spaces Mostow was studying were discrete: in other words, unlike flat tori or two-dimensional hyperbolic manifolds, higher-dimensional locally symmetric spaces could admit only a discrete set of locally symmetric metrics. Mostow has said explicitly that he was motivated by the desire to find a more geometric understanding of this fact.

Another point worth making is that Mostow’s proof is at least as surprising as his theorem. At the time, the study of locally symmetric spaces, or equivalently of semisimple Lie groups and their lattices, was dominated by two sets of techniques: one set that was purely algebraic and another that used classical methods in differential geometry. Mostow’s original proof (which was only for \mathbb{H}^n) uses instead the theory of quasiconformal mappings and some ideas from dynamics. Raghunathan, another leading figure in the field, has said that when he first read Mostow’s paper, he thought it must be by a different man named Mostow. Similar uses of surprising dynamical and analytical ideas to study the same objects occurred almost simultaneously in work of Furstenberg and Margulis. These ideas have had a long and interesting legacy in the study of locally symmetric spaces, semisimple Lie groups, and related objects.

Further Reading

- Furstenberg, H. 1971. Boundaries of Lie groups and discrete subgroups. In *Actes du Congrès International des Mathématiciens, Nice, 1970*, volume 2, pp. 301–6. Paris: Gauthier-Villars.
- Margulis, G. A. 1977. Discrete groups of motions of manifolds of non-positive curvature. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1974*, pp. 33–45. AMS Translations, volume 109. Providence, RI: American Mathematical Society.
- Mostow, G. D. 1973. *Strong Rigidity of Locally Symmetric Spaces*. Annals of Mathematics Studies, number 78. Princeton, NJ: Princeton University Press.
- Thurston, W. P. 1997. *Three-Dimensional Geometry and Topology*, edited by S. Levy, volume 1. Princeton Mathematical Series, number 35. Princeton, NJ: Princeton University Press.

V.27 The \mathcal{P} versus \mathcal{NP} Problem

The \mathcal{P} versus \mathcal{NP} problem is widely considered to be the most important unsolved problem in theoretical computer science, and one of the most important in all of mathematics. \mathcal{P} and \mathcal{NP} are two of the most basic COMPUTATIONAL COMPLEXITY CLASSES [III.10]: \mathcal{P} is the class of all computational tasks that can be performed in a time that is polynomial in the length of the input, and \mathcal{NP} is the class of all computational tasks where a correct answer can be *verified* in a time that is polynomial in the length of the input. An example of the former is multiplying two n -digit integers (which, even if you use long multiplication, takes roughly n^2 arithmetical operations). An example of the latter is searching in a GRAPH [III.34] with n vertices for a set of m vertices, any two of which are joined by an edge: if you are presented with m such vertices, then you just have to check the $\binom{m}{2}$ pairs of those vertices to make sure that each pair is indeed an edge of the graph.

It appears to be *much* harder to find m vertices that are all joined than to check that a given m vertices are all joined. This suggests that problems in \mathcal{NP} are in general harder than problems in \mathcal{P} . The \mathcal{P} versus \mathcal{NP} problem asks for a proof that the complexity classes \mathcal{P} and \mathcal{NP} really are distinct. For a detailed discussion of the problem, see COMPUTATIONAL COMPLEXITY [IV.20].

V.28 The Poincaré Conjecture

The Poincaré conjecture is the statement that a COMPACT [III.9] simply connected smooth n -dimensional

MANIFOLD [I.3 §6.9] must be homeomorphic to the n -sphere S_n . One can think of a compact manifold as a manifold that lives in a finite region of \mathbb{R}^m for some m and that has no boundary: for example, the 2-sphere and the torus are compact manifolds living in \mathbb{R}^3 , while the open unit disk or an infinitely long cylinder is not. (The open unit disk does not have a boundary in an intrinsic sense, but its realization as the set $\{(x, y) : x^2 + y^2 < 1\}$ has the set $\{(x, y) : x^2 + y^2 = 1\}$ as its boundary.) A manifold is called *simply connected* if every loop in the manifold can be continuously contracted to a point. For instance, a sphere of dimension greater than 1 is simply connected but a torus is not (since a loop that “goes around” the torus will always go around the torus, however you continuously deform it). Thus, the Poincaré conjecture asks whether two simple properties of spheres, compactness and simple connectedness, are enough to characterize spheres.

The case $n = 1$ is not interesting: the real line is not compact and a circle is not simply connected, so the hypotheses of the problem cannot be satisfied. POINCARÉ [VI.61] himself solved the problem for $n = 2$ early in the twentieth century, by completely classifying all compact 2-manifolds and noting that in his list of all possible such manifolds only the sphere was simply connected. For a time he believed that he had solved the three-dimensional case as well, but then discovered a counterexample to one of the main assertions of his proof. In 1961, Steven Smale proved the conjecture for $n \geq 5$, and Michael Freedman proved the $n = 4$ case in 1982. That left just the three-dimensional problem open.

Also in 1982, William Thurston put forward his famous *geometrization conjecture*, which was a proposed classification of three-dimensional manifolds. The conjecture asserted that every compact 3-manifold can be cut up into submanifolds that can be given METRICS [III.58] that turn them into one of eight particularly symmetrical geometric structures. Three of these structures are the three-dimensional versions of Euclidean, spherical, and hyperbolic geometry (see SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3 §6]). Another is the infinite “cylinder” $S_2 \times \mathbb{R}$: that is, the product of a 2-sphere with an infinite line. (This is not compact, but that is because the pieces into which one cuts up the manifold may have boundaries that are not included in the pieces.) Similarly, one can take the product of the hyperbolic plane with an infinite line and obtain a fifth structure. The other three are slightly more complicated to describe. Thurston also gave significant evi-

dence for his conjecture by proving it in the case of so-called Haken manifolds.

The geometrization conjecture implies the Poincaré conjecture; both were proved by Grigori Perelman, who completed a program that had been set out by Richard Hamilton. The main idea of this program was to solve the problems by analyzing RICCI FLOW [III.80]. The solution was announced in 2003 and checked carefully by several experts over the next few years. For more details, see DIFFERENTIAL TOPOLOGY [IV.7].

V.29 The Prime Number Theorem and the Riemann Hypothesis

How many prime numbers are there between 1 and n ? A natural first reaction to this question is to define $\pi(n)$ to be the number of prime numbers between 1 and n and to search for a formula for $\pi(n)$. However, the primes do not have any obvious pattern to them and it has become clear that no such formula exists (unless one counts highly artificial formulas that do not actually help one to calculate $\pi(n)$).

The standard reaction of mathematicians to this kind of situation is to look instead for good *estimates*. In other words, we try to find a simply defined function $f(n)$ for which we can prove that $f(n)$ is always a good approximation to $\pi(n)$. The modern form of the prime number theorem was first conjectured by GAUSS [VI.26] (though a closely related conjecture had been made by LEGENDRE [VI.24] a few years earlier). He looked at the numerical evidence, which suggested to him that the “density” of primes near n was about $1/\log n$, in the sense that a randomly chosen integer near n would have a probability of roughly $1/\log n$ of being a prime. This leads to the conjectured approximation of $n/\log n$ for $\pi(n)$, or to the slightly more sophisticated approximation

$$\pi(n) \simeq \int_0^n \frac{dx}{\log x}.$$

The function defined by the integral on the right-hand side is called $\text{li}(n)$ (which stands for the “logarithmic integral” of n). Some care is needed in interpreting the integral because $\log 1 = 0$, but one can avoid this problem by integrating from 2 to n instead, which changes the function by just an additive constant.

The prime number theorem, proved independently by HADAMARD [VI.65] and DE LA VALLÉE POUSSIN [VI.67] in 1896, states that $\text{li}(n)$ is indeed a good approximation to $\pi(n)$, in the sense that the ratio of the two functions tends to 1 as n tends to infinity.

This result is considered one of the great theorems of all time, but it is by no means the end of the story. The proofs of Hadamard and de la Vallée Poussin used the RIEMANN ZETA FUNCTION [IV.2 §3] $\zeta(s)$. The Riemann zeta function is defined to be $1^{-s} + 2^{-s} + 3^{-s} + \dots$ whenever s is a complex number with real part greater than 1; this expression defines a HOLOMORPHIC FUNCTION [I.3 §5.6], which can be extended (by analytic continuation) to a function that is holomorphic on the entire complex plane, except for a pole at 1. This function has zeros, known as “trivial zeros,” at all negative even integers. Riemann proved that the prime number theorem was equivalent to the assertion that the only “nontrivial zeros” were inside the *critical strip*, which consists of those complex numbers with real part strictly between 0 and 1. He also formulated what is often held to be the most important unsolved problem in mathematics, now known as the *Riemann hypothesis*: that in fact the nontrivial zeros all have real part equal to $\frac{1}{2}$. This assertion about the zeros of the zeta function has been shown to be equivalent to a stronger form of the prime number theorem, which states not just that $\pi(n)/\text{li}(n)$ tends to 1, but even that $|\pi(n) - \text{li}(n)| \leq \sqrt{n} \log n$ for every $n \geq 3$. Since $\text{li}(n)$ is around $n/\log n$, which is much bigger than $\sqrt{n} \log n$, this would mean that the error $|\pi(n) - \text{li}(n)|$ was extremely small compared with $\pi(n)$ or $\text{li}(n)$ themselves.

The importance of the Riemann hypothesis goes far beyond its consequences for the distribution of primes: hundreds of statements in number theory have been shown to follow from it. This is particularly true when one considers generalizations of the Riemann hypothesis that apply to a wider class of *L-FUNCTIONS* [III.49]. For example, analogues of the Riemann hypothesis for Dirichlet *L-functions* imply very good estimates for the distribution of primes in arithmetic progressions, from which many further consequences follow.

The prime number theorem and the Riemann hypothesis are discussed in more detail in ANALYTIC NUMBER THEORY [IV.2 §3].

V.30 Problems and Results in Additive Number Theory

Is every even number greater than 4 the sum of two odd primes? Are there infinitely many primes p such that $p + 2$ is also a prime? Is every sufficiently large positive integer the sum of four cubes? These three questions are all famous unsolved problems in number theory: the first is called the *Goldbach conjecture*,

PUP: I confirm that this is the right word here.

the second is the *twin-primes conjecture* (discussed in some detail in ANALYTIC NUMBER THEORY [IV.2]), and the third is a special case of *Waring's problem*, which we shall discuss later.

These three problems belong to an area of mathematics known as *additive number theory*. In order to say in general terms what this area is, it is useful to make some simple definitions. Suppose that A is a set of positive integers. Then the *sumset* of A , denoted $A + A$, is the set of all $x + y$ such that x and y (which are allowed to be equal) both belong to A . For example, if A is the set $\{1, 5, 9, 10, 13\}$, then $A + A$ is the set $\{2, 6, 10, 11, 14, 15, 18, 19, 20, 22, 23, 26\}$. Similarly, the *difference set*, denoted $A - A$, is the set of all $x - y$ such that x and y both belong to A . In the above example, $A - A = \{-12, -9, -8, -5, -4, -3, -1, 0, 1, 3, 4, 5, 8, 9, 12\}$.

Using this language, we can state two of our three problems very succinctly. Let P be the set of all odd primes and let C be the set of all cubes. Then Goldbach's conjecture asks whether $P + P$ is the set $\{6, 8, 10, 12, \dots\}$, and the special case of Waring's problem asks whether every sufficiently large integer belongs to $C + C + C + C$. The twin-primes conjecture is slightly more complicated: it states not just that 2 belongs to the set $P - P$ but that it does so "infinitely many times." (In a similar way, if A is the set in the previous paragraph, then $A - A$ contains the number 4 three times.)

These problems are notoriously difficult. However, remarkably, there are some closely related problems that look just as hard at first, but which have been solved. For instance, *Vinogradov's three-primes theorem* is the statement that every sufficiently large odd integer is the sum of *three* odd primes. Without the "sufficiently large" this would answer the *ternary Goldbach problem*, which asks whether every odd number from 9 onward is a sum of three odd primes. (How large is "sufficiently large"? Well, until recently you needed your number to have about 7 000 000 digits, but in 2002 this was reduced to under 1500 digits.) As for Waring's problem, it is known that every sufficiently large positive integer is a sum of seven cubes. More generally, it seems likely that, for any k , every sufficiently large integer can be written as a sum of at most $100k$ k th powers (where 100 is just a randomly chosen largish number—it is possible that even $4k$ k th powers are enough), and although a proof of this is well beyond today's mathematical technology, it has been shown that a little over $k \log k$ k th powers are enough. Since $\log k$ is a very slowly growing function, this result is,

in a certain sense, not too far from a solution to the problem.

How does one obtain results such as these? Some of the proofs are pretty complicated, so we cannot give a full answer here. However, we can at least explain one idea that is fundamental to many of the arguments, namely the use of *exponential sums*. Let us illustrate it by looking at the beginning of the proof of the Vinogradov three-primes theorem.

Imagine, then, that we have a very large odd integer n and we wish to prove that it is a sum of three odd primes. Here is an argument that strongly suggests that our task is impossible: if n is over three times larger than the largest known prime, as it may very well be, then we cannot produce three primes that add up to n without finding a new prime. Indeed, we could take n to be astronomically large, $10^{10^{100}} + 1$, say, and then $\frac{1}{3}n$ would be far beyond any prime that has ever been discovered or is ever likely to be discovered.

This argument is, however, flawed, and the clue to what is wrong with it lies in the word "produce." We do not have to *produce* the three primes to show that they *exist*, any more than Euclid had to specify an infinite sequence of primes in order to show that there were infinitely many. (For a proof that there are, see [IV.2 §2].) But, one might ask, what alternative could there possibly be to actually finding three odd primes that add up to n ?

This question has a beautifully simple answer: we shall attempt to *count*, or rather *estimate*, the number of triples p_1, p_2, p_3 of odd primes such that $p_1 + p_2 + p_3 = n$. If the estimate we manage to obtain is rather large, and if in addition we can show that it is reasonably accurate, then the actual number of such triples must also be rather large. This will imply that there *is* such a triple, and will not require us to "produce" one.

However, our answer immediately raises a difficult-looking question: how do we estimate the number of such triples? This is where exponential sums come in. We shall use certain properties of the EXPONENTIAL FUNCTION [III.25] to reformulate our counting problem as a problem about estimating a certain integral.

As is customary in this area, let us write $e(x)$ instead of $e^{2\pi i x}$. The two basic properties that we shall use of the function $e(x)$ are that $e(x + y) = e(x)e(y)$ and that $\int_0^1 e(nx) dx = 1$ if $n = 0$, and 0 if n is any other integer. Let us also adopt the convention that if we write $\sum_{p \leq n}$, then we are summing over all odd primes less than or equal to n . Now define a function $F(x)$ by the formula

T&T note: check linebreaks here before CRC is produced as '\allowbreak' added.

$F(x) = \sum_{p \leq N} e(px)$. That is,

$$F(x) = e(3x) + e(5x) + e(7x) + e(11x) + \cdots + e(qx),$$

where q is the largest prime less than or equal to n . This is a sum of exponentials—hence the phrase “exponential sums.” Next, we consider the cube of this function:

$$F(x)^3 = (e(3x) + e(5x) + e(7x) + \cdots + e(qx))^3.$$

When we multiply out the right-hand side, we obtain the sum of all terms of the form $e(p_1x)e(p_2x)e(p_3x)$, where p_1, p_2 , and p_3 are primes between 3 and q .

The integral we shall look at is $\int_0^1 F(x)^3 e(-nx) dx$. From our discussion in the previous paragraph, we know that this will be the sum of all integrals of the form $\int_0^1 e(p_1x)e(p_2x)e(p_3x)e(-nx) dx$. Now the first basic property of $e(x)$ tells us that this last integral is equal to $\int_0^1 e((p_1 + p_2 + p_3 - n)x) dx$, and the second one then tells us that it is 1 if $p_1 + p_2 + p_3 = n$ and 0 otherwise. Therefore, when we sum over all possible triples p_1, p_2, p_3 of odd primes less than or equal to n , we get a contribution of 1 for each triple that adds up to n and 0 for all other triples. In other words, the integral $\int_0^1 F(x)^3 e(-nx) dx$ exactly equals the number of ways of writing n as a sum of three odd primes.

This “reduces” our problem to that of estimating the integral $\int_0^1 F(x)^3 e(-nx) dx$. But the function $F(x)$ looks rather difficult to analyze. Is it really feasible to estimate an expression such as $\sum_{p \leq N} e(px)$, which mixes prime numbers with exponentials?

Surprisingly, it is. The details are complicated, but the fact that it can be done becomes less mysterious after one thinks for a moment about which exponential sums we definitely *can* estimate. Are there at least *some* sets A of integers for which we can handle sums of the form $\sum_{a \in A} e(ax)$? Yes there are: arithmetic progressions. Suppose A is the set $\{s, s + d, s + 2d, \dots, s + (m - 1)d\}$: that is, the arithmetic progression of length m and common difference d that starts at s . Then, using the basic properties of $e(x)$, we find that $\sum_{a \in A} e(ax)$ is

$$\begin{aligned} & e(sx) + e((s + d)x) + \cdots + e((s + (m - 1)d)x) \\ &= e(sx) + e(dx)e(sx) + \cdots + e((m - 1)d)x e(sx) \\ &= e(sx)(1 + e(dx) + e(dx)^2 + \cdots + e(dx)^{m-1}). \end{aligned}$$

This last expression is the sum of a geometric progression that starts at $e(sx)$ and has common ratio $e(dx)$. Using the standard formula and the basic properties of $e(x)$, we deduce that

$$\sum_{a \in A} e(ax) = \frac{e(sx) - e((s + dm)x)}{1 - e(dx)}.$$

Such expressions are useful because they can often be shown to be small. Suppose, for instance, that $|1 - e(dx)|$ is at least as big as some constant c . We know that $|e(sx) - e((s + dm)x)| \leq 2$, so the modulus of the right-hand side is at most $2/c$. If c is not too small, then this shows that there is a huge amount of cancellation in the sum $\sum_{a \in A} e(ax)$: we added together m numbers of modulus 1 and obtained a number of modulus no bigger than $2/c$.

For certain values of x , we can use this simple observation to help us estimate the sum $\sum_{p \in P} e(px)$. What we need to do is express the sum over P as a combination of sums over arithmetic progressions, and this is a very natural thing to do, since P consists of all those integers up to n that do not lie in certain arithmetic progressions (such as 14, 21, 28, 35, 42, ...). So we can begin by taking the sum $\sum_{t=1}^n e(tx)$. From this we need to subtract the contribution from all even integers, which is $\sum_{t \leq n/2} e(2tx)$. We also need to subtract the contribution from multiples of 3, apart from 3 itself. This contribution is $\sum_{1 < t \leq n/3} e(3tx)$. Now we find that we have subtracted the contribution from multiples of 6 twice, so we correct for that by adding $\sum_{t \leq n/6} e(6tx)$.

This process can be continued, and it leads to a way of decomposing the sum over primes into a combination of sums over geometric progressions. If x is not close to a rational with small denominator, then most of the common ratios are far from 1, so most of the sums over progressions are small. Unfortunately, there are too many of them for this simple argument to lead to a useful estimate. However, there is a more sophisticated argument with a similar flavor that does.

What happens if x is close to a rational with small denominator? For example, what can we say about the sum $\sum_{p \leq n} e(p/3)$? Here we use more direct methods: it is known that roughly half of all primes are 1 (mod 3) and half are 2 (mod 3) (see [IV.2 §4]), which tells us that this sum is roughly $(|P|/2)(e(p/3) + e(2p/3))$, where $|P|$ denotes the size of the set P .

For very similar reasons, in Waring’s problem one finds oneself wanting to know about exponential sums such as $G(x) = \sum_{t=0}^m e(t^k x)$. Again, one can sometimes estimate these by reducing them to sums of geometric progressions. This is easiest to show in the case $k = 2$. The idea is to look not at $G(x)$ directly but at $|G(x)|^2$, which a moment’s calculation shows is equal to $\sum_{t=0}^m \sum_{u=0}^m e((t^2 - u^2)x)$. Now $t^2 - u^2 = (t + u)(t - u)$, so we can change variables, setting $v = t + u$ and $w = t - u$. This gives us the sum $\sum_{(v,w) \in V} e(vwx)$, where V is the set of all (v, w) such that $(v + w)/2$

T&T note: check linebreaks here before CRC is produced as ‘\allowbreak’ added.

and $(v - w)/2$ (which equal t and u , respectively) are both between 0 and m . For each v the set of possible values of w is an arithmetic progression, so we have decomposed $|G(x)|^2$ into a sum of sums of geometric progressions, one for each v .

So far we have been looking at so-called *direct* problems in additive number theory. These are problems where one specifies a set and then tries to understand its sumset or difference set. We have only scratched the surface of the subject: other related results and techniques are discussed in [IV.2] (see in particular sections 7, 9, and 11).

Direct problems have a long history, but in recent years another class of problems, called *inverse* problems, have become an important focus of research as well. These concern the following broad question: if you are given information about a sumset or a difference set, what can you deduce about the original set? We end by describing one of the highlights of this kind of additive number theory, called *Freiman's theorem*.

It is not hard to prove that if A is any set of integers of size n , then the size of $A + A$ must be between $2n - 1$ and $n(n + 1)/2$. (The first happens if A is an arithmetic progression and the second happens if all the sums you can make are different.) What can we say about A if the size of $A + A$ is at most $100n$, or, more generally, is at most Cn for some constant C that remains fixed as n tends to infinity?

Suppose that we can find an arithmetic progression P of size at most $50n$ such that A is a subset of P . Then $A + A$ is a subset of $P + P$, which has size $100n - 1$. So if A is two percent of an arithmetic progression, then $A + A$ has size at most $100n$. However, there are other ways of producing such sets. Suppose, for instance, that A consists of all numbers of up to seven digits such that the third, fourth, and fifth digits from the end are 0: that is, numbers such as 35 000 26 or 99 000 90. There are $100 \times 100 = 10\,000$ of these. If we add two of them together, then we get a number like 138 001 62 or 141 000 68, which is made up of a number between 0 and 198, followed by two 0s, followed by a second number between 0 and 198 (written with 0s in front if these are needed to make it up to three digits). There are 199×199 of these, which is less than 40 000. Therefore, the size of $A + A$ is less than four times the size of A . However, A does not fill up two percent of any arithmetic progression P : such a progression would have to have common difference 1 and include both the numbers 0 and 99 000 99, and 10 000 is nothing like two percent of 99 000 100.

However, A is a very structured set: it is an example of a *two-dimensional* arithmetic progression. Roughly speaking, an ordinary, or one-dimensional, arithmetic progression is one that you build up by starting with a number s and repeatedly adding another one, d , called the common difference. You build up a *two-dimensional* arithmetic progression by using *two* “common differences” d_1 and d_2 . That is, you have a starting number s and you look at numbers of the form $s + ad_1 + bd_2$, specifying that a should be between 0 and $m_1 - 1$ and b should be between 0 and $m_2 - 1$. Our set A is a two-dimensional progression with $s = 0$, $d_1 = 1$, $d_2 = 100\,000$, and $m_1 = m_2 = 100$.

In a similar way one can define higher-dimensional progressions. It is not hard to show that if P is an r -dimensional progression, then the size of $P + P$ is less than 2^r times the size of P . Therefore, if A is a subset of P and the size of P is at most C times the size of A , then the size of $A + A$ is at most the size of $P + P$, which is at most $2^r C$ times the size of A .

This tells us that if A is a large subset of a low-dimensional arithmetic progression, then A has a small sumset. Freiman's theorem is the remarkable statement that these are the *only* sets with small sumsets. That is, if $A + A$ is not much larger than A , then there must be some low-dimensional arithmetic progression P that contains A and is not much bigger than A . Exponential sums are vital for the proof of this theorem as well. Freiman's theorem has had many applications, and is likely to have many more.

V.31 From Quadratic Reciprocity to Class Field Theory

Kiran S. Kedlaya

The law of quadratic reciprocity, discovered by EULER [VI.19] and first proved by GAUSS [VI.26] (who dubbed it his *theorema aureum*, or golden theorem), is considered a crown jewel of number theory, and with good cause. Whereas its statement could be rediscovered by a sufficiently ingenious student (indeed, it actually has been rediscovered on a regular basis at the Arnold Ross mathematics summer program for several decades), rare is the student who comes up with a proof unassisted.

The law is most conveniently stated in a formulation due to LEGENDRE [VI.24]. For n an integer not divisible by the prime p , write $(\frac{n}{p}) = 1$ if n is congruent to some perfect square modulo p , and $(\frac{n}{p}) = -1$ if it is not. Then

quadratic reciprocity states the following. (The prime 2 must be treated separately.)

Theorem (quadratic reciprocity). *Suppose that p and q are two different primes, neither equal to 2. Then $\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = -1$ if p and q are both congruent to 3 modulo 4, and $\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = 1$ otherwise.*

For instance, if $p = 13$ and $q = 29$, then $\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = 1$. Since 29 is congruent modulo 13 to the perfect square 16, it must be that 13 is congruent to some perfect square modulo 29, and in fact $100 = 3 \cdot 29 + 13$.

This statement is simple but also mysterious, because it violates our intuition that congruences modulo different primes should act independently. For instance, the Chinese remainder theorem asserts that (in a suitably precise sense) knowing that a random integer is odd or even does not prejudice it toward having any particular remainder modulo 3. Number theorists are fond of using geometric language to describe this situation, referring to phenomena associated with congruences modulo a single prime (or a power of a single prime) as *local* phenomena (see LOCAL AND GLOBAL IN NUMBER THEORY [III.53]). The Chinese remainder theorem can be interpreted as saying that local phenomena at one point really are local, in that they do not influence local phenomena at another point. However, just as a particle physicist cannot explain the behavior of the universe by analyzing individual particles in isolation, one cannot hope to understand the behavior of integers by looking at individual primes in isolation. Quadratic reciprocity thus emerges as one of the first known examples of a *global* phenomenon, proving to be a “fundamental force” that binds together two different primes. The interplay between local and global is built thoroughly into our modern understanding of number theory, but the phenomenon of quadratic reciprocity was where it first came to light.

Another indication of the fundamental nature of quadratic reciprocity is that it admits proofs using many different techniques. Gauss himself devised eight proofs in his lifetime, and nowadays dozens of proofs are available. These suggest numerous directions of generalization; here we will focus on the direction that led historically to class field theory. Among the many fascinating sidelights that this will force us to omit is the theory of Gauss sums and its surprisingly diverse range of applications, such as Kolyagin’s work on THE BIRCH-SWINNERTON-DYER CONJECTURE [V.4], and the

use of number theory in CRYPTOGRAPHY [VII.7] and other areas of computer science.

Euler had sought reciprocity laws for perfect third and fourth powers, but had had limited success. Gauss succeeded in formulating such laws (but not proving them; that fell to Eisenstein later) by realizing that one could only properly understand them by stepping out of the ring of integers.

Let us see this explicitly for fourth powers. Let p and q be primes that are both congruent to 1 modulo 4. The reciprocity between p being congruent to a fourth power modulo q and vice versa cannot be easily stated in terms of p and q . Instead, we must recall a result of FERMAT [VI.12]: we can write $p = a^2 + b^2$ and $q = c^2 + d^2$, where each of the pairs (a, b) and (c, d) is unique up to changing signs and ordering. In other words, in the ring of complex numbers whose real and imaginary parts are integers (now called the *Gaussian integers*), we have $p = (a + bi)(a - bi)$ and $q = (c + di)(c - di)$.

Gauss defined an analogue of the Legendre symbol as follows. It was already known to Euler that

$$\left(\frac{n}{p}\right) \equiv n^{(p-1)/2} \pmod{p};$$

to see that the right-hand side is either 1 or -1 , note that it squares to 1 by FERMAT’S LITTLE THEOREM [III.60], and the equation $x^2 = 1$ has just these two roots. Gauss similarly defined

$$\left(\frac{c + di}{a + bi}\right)_4$$

to be i^k , for the unique choice of k modulo 4 for which $i^k \equiv (c + di)^{(a^2 + b^2 - 1)/4} = (c + di)^{(p-1)/4} \pmod{a + bi}$.

Here we say that two integers are congruent mod $a + bi$ if their difference is a multiple of $a + bi$ by a Gaussian integer. The existence of such k again follows from Fermat’s little theorem: if you expand $(c + di)^p$, then all the binomial coefficients are multiples of p apart from the first and the last, so you obtain $c^p + (di)^p$, which equals $c + di$ by Fermat’s theorem and the assumption that p is congruent to 1 mod 4; it follows that $(c + di)^{p-1} \equiv 1$. (Alternatively, one can prove this by showing that the Gaussian integers mod $a + bi$ form a group of order $p - 1$ and applying Lagrange’s theorem.)

Before stating the reciprocity law, we must stamp out the ambiguity in the choice of a, b, c , and d . We require that a and c must be odd, and that $a + b - 1$ and $c + d - 1$ must be divisible by 4. (Note that we can still flip the signs of b and d .)

Theorem (quartic reciprocity). *With p, q, a, b, c , and d as above, we have*

$$\left(\frac{a+bi}{c+di}\right)_4 \left(\frac{c+di}{a+bi}\right)_4 = -1$$

if p and q are both congruent to 5 modulo 8, and

$$\left(\frac{a+bi}{c+di}\right)_4 \left(\frac{c+di}{a+bi}\right)_4 = 1$$

otherwise.

One might expect to find an n th power reciprocity law that looks like this by working with the ring generated by a primitive n th root of 1. What complicates matters is that this ring does not enjoy the UNIQUE FACTORIZATION PROPERTY [IV.1 §§4–8] (whereas the usual integers and the Gaussian integers both do). This was remedied only by KUMMER's [VI.40] theory of IDEALS [III.83 §2] (short for “ideal numbers”). An ideal is a set that has the typical properties of the set of all multiples of a given number, but it can be more general. (Even if an ideal is the set of all multiples of some number, that number is not unique, since one can multiply it by a unit. For instance, both 2 and -2 generate the ideal of all even numbers.) Using Kummer's theory, Kummer and Eisenstein managed to formulate broad generalizations of quadratic reciprocity for higher powers.

HILBERT [VI.63] then realized that these should fit together as part of some sort of maximally general reciprocity law. He also gave a candidate for this law, inspired by a reformulation of quadratic reciprocity itself in terms of the *norm residue symbol*. For a prime p , and any nonzero integers m and n , the norm residue symbol $(\frac{m,n}{p})$ equals 1 if, for all sufficiently large k , the equations $mx^2 + ny^2 \equiv z^2 \pmod{p^k}$ have solutions where x, y , and z are not all divisible by p^k ; otherwise the symbol equals -1 . In other words, the symbol equals 1 if the equation $mx^2 + ny^2 = z^2$ has a solution in the p -ADIC NUMBERS [III.53].

Hilbert's formulation of quadratic reciprocity is that, for any nonzero m and n ,

$$\prod_p \left(\frac{m,n}{p}\right) = 1,$$

where the product is taken over all primes p and the prime $p = \infty$. The latter requires some explanation: we write $(\frac{m,n}{\infty}) = 1$ if and only if m and n are not both negative, i.e., if the equation $mx^2 + ny^2 = z^2$ has a solution in the *real* numbers. This fits into a general pattern, that conditions quantified over “all prime numbers” must also account for the so-called infinite prime.

It should also be clarified that Hilbert's product only makes sense by virtue of the fact that, for fixed m and n , $(\frac{m,n}{p}) = 1$ for all but finitely many p . This is because in general, since approximately half the integers mod p^k are quadratic residues, it is easy to solve the equation $mx^2 + ny^2 = z^2$: difficulties arise only when multiplication by m or n identifies many of these quadratic residues. For instance, if m and n are (positive) prime numbers, then only those two primes contribute to the product; the two resulting factors can be related to $(\frac{m}{n})$ and $(\frac{n}{m})$, which leads back to quadratic reciprocity.

Using this formulation, Hilbert was able to state and prove a form of quadratic reciprocity over any NUMBER FIELD [III.65], in which the corresponding product of symbols is quantified over the prime ideals of the number field (together with some “infinite primes”). Hilbert also conjectured a higher-power reciprocity law over any number field. That conjecture was tackled by Hasse, Takagi, and finally ARTIN [VI.86], who stated a general reciprocity law. Its statement is a bit too technical to include here; we limit ourselves to observing that Artin's reciprocity law, when applied to a number field K , describes certain norm residue symbols in terms of Abelian extensions of K , i.e., number fields containing K whose groups of symmetries (GALOIS GROUPS [V.24]) are commutative.

The Abelian extensions of \mathbb{Q} are easy to describe: the Kronecker-Weber theorem asserts that they are all contained in fields generated by roots of 1. This explains the role of the roots of 1 in the classical reciprocity laws. However, describing the Abelian extensions of an arbitrary number field K is somewhat harder. They can at least be classified in terms of the structure of the field K itself; this is what is commonly referred to as *class field theory*.

However, the problem of explicitly specifying generators of the Abelian extensions of K (Hilbert's twelfth problem) remains mostly unsolved, except in some special cases. For instance, the theory of ELLIPTIC FUNCTIONS [V.34] solves this problem for fields of the form $\mathbb{Q}(\sqrt{-d})$ with $d > 0$ via the theory of *complex multiplication*. Some additional examples emerged from the work of Shimura on MODULAR FORMS [III.61], leading to the *Shimura reciprocity law*.

This last example shows that the story of reciprocity laws is not yet complete. Any new instance of explicit class field theory would reveal another reciprocity law that had previously been hidden from view. Some exciting new conjectures in this direction have

been advanced by Bertolini, Darmon, and Dasgupta, who have proposed some new constructions of Abelian extensions using p -adic analysis. These are analogous to the aforementioned constructions using elliptic functions, in which one evaluates a transcendental function at a special value. At first, there seems to be no reason to expect the resulting complex number to have any special properties, but in fact it turns out to be an algebraic number that generates an appropriate Abelian extension of the base field. While one can check in individual examples, using computer calculations, that the construction seems to be converging p -adically to a particular generator of the right field, a proof seems out of reach at present.

Further Reading

Ireland, K., and M. Rosen. 1990. *A Classical Introduction to Modern Number Theory*, 2nd edn. New York: Springer.
 Lemmermeyer, F. 2000. *Reciprocity Laws, from Euler to Eisenstein*. Berlin: Springer.

V.32 Rational Points on Curves and the Mordell Conjecture

Suppose that we wish to study a Diophantine equation such as $x^3 + y^3 = z^3$. A simple observation we can make is that studying integer solutions to this equation is more or less equivalent to studying rational solutions to the equation $a^3 + b^3 = 1$: indeed, if we had integers x , y , and z such that $x^3 + y^3 = z^3$, then we could set $a = x/z$ and $b = y/z$ and obtain rational numbers with $a^3 + b^3 = 1$. Conversely, given rational numbers a and b with $a^3 + b^3 = 1$, we could multiply a and b by the lowest common multiple z of their denominators and set $x = az$ and $y = bz$, obtaining integers x , y , and z such that $x^3 + y^3 = z^3$.

The advantage of doing this is that it reduces the number of variables by 1 and focuses our attention on the plane curve $u^3 + v^3 = 1$, which is a simpler object than the surface $x^3 + y^3 = z^3$. A curve of this kind, defined by one or more polynomial equations, is called an *algebraic curve*.

Even though we are interested in rational points on the curve, it can be helpful to regard the curve as an abstract object that has many manifestations. (See ARITHMETIC GEOMETRY [IV.5] for a fuller discussion of this point.) For instance, if we think of u and v as complex numbers, then the “curve” $u^3 + v^3 = 1$ becomes a two-dimensional object, which means that it starts

to have a genuinely interesting geometry. To be precise, it can be regarded as a two-dimensional MANIFOLD [I.3 §6.9] living in \mathbb{R}^4 . From a *complex* perspective it is a one-dimensional subset of \mathbb{C}^2 , but from either perspective it has a potentially interesting topology. For instance, if we COMPACTIFY [III.9] the curve by considering it as a subset not of \mathbb{C}^2 but of the complex PROJECTIVE PLANE [I.3 §6.7], then we turn it into a compact surface. As such, it must have a GENUS [III.33], which, roughly speaking, tells us how many holes it has.

Surprisingly, it turns out that this geometrical definition of the genus of a curve is intimately related to the algebraic question of how many rational points the curve contains. Consider, for instance, the curve $u^2 + v^2 = 1$, which corresponds to the Diophantine equation $x^2 + y^2 = z^2$. Since there are infinitely many genuinely Pythagorean triples that are not multiples of each other, there are infinitely many rational points on the curve $u^2 + v^2 = 1$. In order to calculate the genus of the curve, we first rewrite it as $(u + iv)(u - iv) = 1$. This shows that the function $(u, v) \mapsto u + iv$ is a homeomorphism from the curve to the set $\mathbb{C} \setminus \{0\}$ of all nonzero complex numbers, which itself is homeomorphic to a sphere with two points removed. The compactification adds in these points, giving us a surface of genus 0, so we say that the curve $u^2 + v^2 = 1$ has genus 0. It turns out that a curve of genus 0 always has either no rational points or infinitely many.

In general, the larger the genus, the harder it is to find rational solutions. A curve of genus 1 is called an ELLIPTIC CURVE [III.21]. It is possible for an elliptic curve to contain infinitely many rational points as well, but the set of such points turns out to have a very restricted structure. To explain this, let us consider an elliptic curve E of the form $y^2 = ax^3 + bx^2 + cx + d$ (a form into which any elliptic curve can be put). If we think of it as a curve in \mathbb{R}^2 , then we can define a binary operation on it as follows: given any two points P and Q on E , let L be the line through P and Q (where we define this to be the tangent to the curve at P if $P = Q$). In general, L intersects E in three points, of which P and Q are two; let R' be the third. Finally, let R be the reflection of R' in the x -axis (which also belongs to E because E has the form $y^2 = f(x)$). This construction of R from P and Q , which is illustrated in figure 1, defines a binary operation on the points of E . Remarkably, this binary operation turns E into an Abelian group, at least when we also include a point at infinity and adopt the convention that the point at infinity is the intersection of E with any vertical line. The point at infinity is the iden-

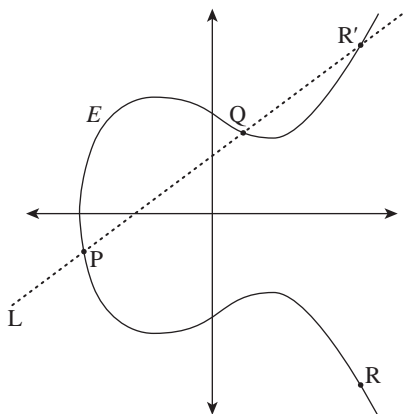


Figure 1 The group law for an elliptic curve.

tity of the group, since a vertical line through a point P intersects E in the reflection P' of P in the x -axis, and when we reflect P' in the x -axis we get P again.

It is laborious, but basically straightforward, to come up with a formula for the “group law” of an elliptic curve—that is, a formula for the coordinates of R in terms of the coordinates of P and Q . Once one does so, it becomes clear that if P and Q have rational coordinates, then so does R . Thus, the set of all rational points on an elliptic curve E forms a subgroup. This simple fact can be used to produce rather easily some very large solutions to the corresponding Diophantine equations. For instance, one can start with a small solution, associate with it a rational point P , and then use the formula for the binary operation to calculate $2P$, then $4P$, then $8P$, and so on. Unless $nP = 0$ for some n (which can certainly happen), in no time at all one has a point on the curve with rational coordinates that have huge numerators and denominators. To give an idea of the sort of solutions that can be obtained in this way, take the elliptic curve $y^2 = x^3 - 5x$ and let P be the point $(-1, 2)$ (which lies on the curve since $2^2 = (-1)^3 - 5(-1)$). If you calculate $5P$ using the group law, then you obtain the point $(-5\,248\,681/4\,020\,025, 16\,718\,705\,378/8\,060\,150\,125)$. In general, the number of digits needed to express the point nP grows exponentially with n .

In the early twentieth century, POINCARÉ [VI.61] conjectured that the subgroup of rational points on an elliptic curve was finitely generated. This conjecture was proved by Louis Mordell in 1922. Thus, although a curve of genus 1 may have infinitely many rational points, there is a finite set of these points that can be used to build up all the others: this is the sense in

which the structure of the set of rational solutions is restricted.

Mordell conjectured that a curve of genus at least 2 could contain only finitely many points. This was a remarkable conjecture: if true, it would apply to an extremely wide class of Diophantine equations, proving that all of them had at most finitely many solutions (up to a multiple). Just one of its many implications was that for each $n \geq 3$ the Fermat equation $x^n + y^n = z^n$ had at most finitely many solutions with x , y , and z coprime. However, it is one thing to make a very general conjecture and quite another to prove it, and for a long time the consensus was that the Mordell conjecture, like many other conjectures in number theory, was way beyond what anybody could prove. It therefore came as a big surprise when Gerd Faltings proved the conjecture in 1983.

As a result of Faltings’s proof, our knowledge about Diophantine equations took a huge leap forward. The theorem has subsequently been given a variety of different proofs, some of them simpler than that of Faltings. However, remarkable as these proofs are, they do have some limitations. One is that they are *ineffective*. That is, even though Faltings’s theorem tells us that certain curves have finitely many rational points, no known proof gives any bound on the sizes of the numerators and denominators of the coordinates of those points, so we do not have any way of knowing whether we have found all of them. This aspect of the theorem is common in number theory: another example of a famous theorem that is ineffective is ROTH’S THEOREM [V.25]. To find effective versions of these theorems would be a further remarkable breakthrough. (Variants of THE ABC CONJECTURE [V.1] would imply effective versions of these results, but the ABC conjecture seems even further out of reach now than Mordell’s conjecture seemed before Faltings proved it.)

At the beginning of this article, we simplified the equation $x^3 + y^3 = z^3$ so that we were looking at a curve rather than a surface. But we obviously cannot always do that. For instance, if we apply the same procedure to the equation $x^5 + y^5 + z^5 = w^5$, then we obtain the two-dimensional surface $t^5 + u^5 + v^5 = 1$. Our knowledge about rational points on varieties (that is, sets defined by polynomial equations) of dimension greater than 1 is very limited. However, there is at least a definition of a “variety of general type” that serves as an analogue of the notion of a curve of genus at least 2. One cannot expect such a variety to contain only finitely many rational points, but a higher-dimensional

analogue of the Mordell conjecture, due to Serge Lang, asserts that the rational points on a variety X of general type must all be contained in a union of finitely many lower-dimensional subvarieties of X . This conjecture is considered to be well out of reach of present methods: indeed, it is not even universally believed.

V.33 The Resolution of Singularities

Virtually all important mathematical structures come with a notion of equivalence. For instance, we regard two GROUPS [I.3 §2.1] as equivalent if they are ISOMORPHIC [I.3 §4.1], and we regard two TOPOLOGICAL SPACES [III.92] as equivalent if there is a continuous map from one to the other with a continuous inverse (in which case we say that they are *homeomorphic*). In general, a notion of equivalence is useful if properties that we are interested in are unaffected when we replace an object by an equivalent one: for example, if G is a finitely generated Abelian group and H is isomorphic to G , then H is a finitely generated Abelian group.

A useful notion of equivalence for ALGEBRAIC VARIETIES [IV.4 §7] is that of *birational* equivalence. Roughly speaking, two varieties V and W are said to be birationally equivalent if there is a rational map from V to W with a rational inverse. If V and W are presented as solution sets of equations in some coordinate system, then these rational maps are just rational functions in the coordinates that send points of V to points of W . However, it is important to understand that a rational map from V to W is not literally a function from V to W , because it is allowed to be undefined at certain points of V .

Consider, for example, how we might map the infinite cylinder $\{(x, y, z) : x^2 + y^2 = 1\}$ to the cone $\{(x, y, z) : x^2 + y^2 = z^2\}$. An obvious map would be the function $f(x, y, z) = (zx, zy, z)$, which we could try to invert using the map $g(x, y, z) = (x/z, y/z, z)$. However, g is not defined at the point $(0, 0, 0)$. Nevertheless, the cylinder and the cone are birationally equivalent, and algebraic geometers would say that g “blows up” the point $(0, 0, 0)$ to the circle $\{(x, y, z) : x^2 + y^2 = 1, z = 0\}$.

The main property of a variety V that is preserved by birational equivalence is the so-called *function field* of V , which consists of all rational functions defined on V . (What precisely this means is not completely obvious: in some contexts, V is a subset of a larger space such as \mathbb{C}^n in which one can talk about ratios of polynomials, and then one possible definition of a rational function on V is that it is an equivalence class of such ratios,

where two of them are counted as equivalent if they take the same values on V . See ARITHMETIC GEOMETRY [IV.5 §3.2] and QUANTUM GROUPS [III.77 §1] for further discussion of this equivalence relation.)

A famous theorem of Hironaka, proved in 1964, states that every algebraic variety (over a field of characteristic 0) is birationally equivalent to an algebraic variety without singularities, with some technical conditions on the birational equivalence that are needed for the theorem to be interesting and useful. The example given earlier is a simple illustration: the cone has a singularity at $(0, 0, 0)$ but the cylinder is smooth everywhere. Hironaka’s proof was well over two hundred pages long, but his argument has since been substantially simplified by several authors.

For a further discussion of the resolution of singularities, see ALGEBRAIC GEOMETRY [IV.4 §9].

The Riemann Hypothesis

See THE PRIME NUMBER THEOREM AND THE RIEMANN HYPOTHESIS [V.29]

V.34 The Riemann–Roch Theorem

A RIEMANN SURFACE [III.81] is a MANIFOLD [I.3 §6.9] that “looks locally like \mathbb{C} ,” in the usual sense of this sort of phrase. In other words, every point has a neighborhood that can be mapped bijectively to an open subset of \mathbb{C} , and where two such neighborhoods overlap, the “transition functions” are HOLOMORPHIC [I.3 §5.6]. One can think of a Riemann surface as the most general sort of set on which the notion of a holomorphic function (that is, a complex-differentiable function) of one complex variable makes sense.

The definition of differentiability is a local one: a function is differentiable if and only if a certain condition holds at each point z , and the condition at z depends only on the behavior of f at points very close to z . However, one of the surprises of complex analysis is that holomorphic functions are much more global than their basic definition would lead one to expect. Indeed, if you know the values of a holomorphic function $f : \mathbb{C} \rightarrow \mathbb{C}$ at every point in a small neighborhood of a single point z , then you can deduce its values at every point in \mathbb{C} . And the same is true if you replace \mathbb{C} by any other (connected) Riemann surface.

Here is a second illustration of the global nature of holomorphic functions. One of the most basic Riemann surfaces is the so-called *Riemann sphere* $\hat{\mathbb{C}}$, which is

obtained from \mathbb{C} by adding a “point at infinity.” A function $f : \hat{\mathbb{C}} \rightarrow \mathbb{C}$ is said to be holomorphic if the following conditions hold:

- f is differentiable at every point of \mathbb{C} ;
- $f(z)$ tends to a limit w as $z \rightarrow \infty$ in any direction;
- w is the value of f at ∞ .

What, then, are the holomorphic functions from $\hat{\mathbb{C}}$ to \mathbb{C} ? A holomorphic function f is continuous, from which it follows that if $f(z)$ tends to a limit as $z \rightarrow \infty$, then f is bounded on \mathbb{C} . But a well-known theorem of LIOUVILLE [VI.39] states that a bounded holomorphic function defined on all of \mathbb{C} must be constant. So the only holomorphic functions from $\hat{\mathbb{C}}$ to \mathbb{C} are constant!

One might take the attitude that it was slightly artificial to consider maps from $\hat{\mathbb{C}}$ to \mathbb{C} . Why not look at maps from $\hat{\mathbb{C}}$ to $\hat{\mathbb{C}}$? Such maps are equivalent to functions from \mathbb{C} to \mathbb{C} that are allowed to tend to infinity at a finite set of points z_1, \dots, z_k , called *poles*, and must tend to a limit as $z \rightarrow \infty$. (This limit is allowed to be the point ∞ . We say that $f(z) \rightarrow \infty$ as $z \rightarrow \infty$ if we can make $|f(z)|$ arbitrarily large by making $|z|$ large enough. Note that some familiar functions such as e^z are ruled out since it is possible for $|z|$ to be large and e^z to be small.) Functions with this property are called *meromorphic*. A typical example is z , or z^2 , or $(1+z)/(1-z)$, or indeed any rational function in z ; it can in fact be shown that any meromorphic function from $\hat{\mathbb{C}}$ to $\hat{\mathbb{C}}$ is rational.

The notion of a meromorphic function also makes sense on other Riemann surfaces. One can think of it as a function that is holomorphic except at a set of isolated points where it tends to infinity. (If the function is defined on \mathbb{C} , there may be infinitely many such points, but a COMPACT [III.9] surface such as $\hat{\mathbb{C}}$ cannot contain infinitely many points that are all isolated from each other, so a meromorphic function on a compact surface has at most finitely many poles.)

A particularly important example is when the Riemann surface in question is a torus. We can regard such a surface as the QUOTIENT [I.3 §3.3] of \mathbb{C} by the lattice generated by two complex numbers u and v such that u/v is not real. There is then a one-to-one correspondence between functions defined on the torus and functions f defined on \mathbb{C} that are *doubly periodic*, in the sense that $f(z+u)$ and $f(z+v)$ are both equal to $f(z)$ for every z . Liouville’s theorem again implies that if such a function is holomorphic then it is constant; however, there are interesting examples of doubly peri-

odic *meromorphic* functions. Such functions are called *elliptic functions*.

Even here, the global nature, or “rigidity,” of holomorphic functions asserts itself, by greatly restricting the supply of elliptic functions. Indeed, one can define a single function, called the *Weierstrass P-function* \wp , with the property that any other elliptic function with respect to a given pair of generators u and v can be expressed as a rational function of \wp and its derivative. Weierstrass’s function (for the generators u and v) is given by the formula

$$\wp(z) = \frac{1}{z^2} + \sum_{(n,m) \neq (0,0)} \left(\frac{1}{(z - mu - nv)^2} - \frac{1}{(mu + nv)^2} \right).$$

Notice that the double periodicity is built into the definition, and that \wp has a pole at every point in the lattice generated by u and v . If we think of \wp as a function on the torus, then it has just one pole. Near this pole, f tends to infinity at the same rate as the function $1/z^2$ does when z tends to 0; we say that the pole has *order 2*. More generally, if f tends to infinity at the same rate as $1/z^k$, then the pole has *order k*.

Suppose we take a compact Riemann surface S and choose from it a finite set of points z_1, \dots, z_r . Given a sequence d_1, \dots, d_r of positive integers, can we find a meromorphic function f defined on S such that its poles are z_1, \dots, z_r and such that for each i the order of the pole at z_i is at most d_i ? The results mentioned so far would lead us to expect that this might be possible, but that there would probably not be a huge supply of such functions. Since a linear combination of such functions gives us another one, the set of functions we are interested in forms a VECTOR SPACE [I.3 §2.3], so we could hope to quantify “how many” functions there are by investigating the dimension of this space.

As we might by now expect, this dimension turns out to be finite. RIEMANN [VI.49] proved that if the poles are required to be *simple* (that is, $d_i = 1$ for $i = 1, 2, \dots, r$), then the dimension l is at least $r - g + 1$, where g is the GENUS [III.33] of the surface, which means, roughly speaking, the number of holes it has. This result is called *Riemann’s inequality*. Roch’s contribution was to interpret the difference between l and $r - g + 1$ as the dimension of another space of functions. This often makes it possible to calculate the dimension l exactly. For instance, under certain circumstances one can show that the dimension of the space of functions identified by Roch is 0, in which case $l = r - g + 1$. In particular, this is the case when $r \geq 2g - 1$.

The original question we asked was more general in that we did not require the poles to be simple: rather,

we wanted the order of the pole at z_i to be at most d_i . However, the result generalizes straightforwardly, and l is now at least $d_1 + \cdots + d_r - g + 1$, with the difference again equal to the dimension of a certain space of functions that one can define. One can even ask for some of the d_i to be negative, interpreting a “pole of order at most d_i ” to mean a zero of multiplicity at least $-d_i$.

The Riemann–Roch theorem is a basic tool for computing the dimensions of spaces of holomorphic or meromorphic functions on compact surfaces (which is often equivalent to requiring them to obey certain symmetry conditions). Let us begin with a very simple example. It is not hard to show that every meromorphic function defined on the Riemann sphere with at most simple poles at 0 and 1 has to take the form $a + b/z + c/(z - 1)$. This is a three-dimensional space, and that is what the Riemann–Roch theorem predicts. A more sophisticated example concerns the Weierstrass P -function. We saw earlier that this is a doubly periodic meromorphic function defined on \mathbb{C} with a pole of order 2 at each point in the lattice generated by u and v . The existence (and essential uniqueness) of such a function can be proved more abstractly with the help of the Riemann–Roch theorem: it shows that the space of such functions has dimension 2, so they can all be built out of a single function \wp and the constant functions. Similarly, the theorem can be used to compute dimensions of spaces of MODULAR FORMS [III.61].

The Riemann–Roch theorem has been reformulated and generalized many times, which has made it even more useful as a computational tool, and a central result in algebraic geometry: for example, Hirzebruch found a higher-dimensional generalization, which was generalized further by Grothendieck to a statement about advanced concepts in modern algebraic geometry such as SCHEMES [IV.5 §3] and “sheaves.” Hirzebruch’s generalization, like the classical result about curves, expresses an analytically defined quantity in terms of purely topological invariants: it is this feature of both results that underlies their importance. Another generalization of which the same can be said is the famous ATIYAH–SINGER INDEX THEOREM [V.2], which has itself been generalized several times.

V.35 The Robertson–Seymour Theorem

Bruce Reed

A *graph* G is a mathematical structure that consists of a set $V(G)$ of *vertices* and a set $E(G)$ of *edges* that link

some of the vertices. Graphs can be used to represent many different networks in an abstract way. For example, the vertices might represent cities, and the edges might represent highways linking the cities; similarly, we could use a graph to represent which islands of an archipelago are linked by bridges, or to represent the wires of a telephone network. Among graphs there are certain families of “nice” graphs. One such family is the family of *cycles*: a k -cycle is a set of k vertices arranged around a circle with each point joined by an edge to the points immediately before and after it. Another family is that of *complete graphs*: the complete graph of order k consists of k vertices, all pairs of which are joined.

An important concept in graph theory, particularly when families of graphs are involved, is that of a *minor*. Given a graph G , a minor of G is any graph you can obtain by applying a sequence of operations of two kinds, known as contractions and deletions, applied to edges. To *contract* the edge that joins two vertices x and y , one “fuses” x and y into a single vertex, joining it to all the vertices that were previously joined to either x or y . For example, if you contract an edge of a 9-cycle, you will obtain an 8-cycle. *Deleting* an edge means what one would guess: for example, if you delete an edge from a 9-cycle you will get a *path* with nine vertices and eight edges.

It is not hard to check that a graph H is a minor of G if and only if we can find a collection of disjoint subsets of G , one for each vertex of H , with the following properties: they should be *connected*, which means that any two vertices in one of the subsets are joined to each other by a path in that subset, and for any pair of vertices in H that are linked by an edge in H the two corresponding subsets of G should be linked by an edge. For example, a graph has a 3-cycle (or *triangle*) as a minor if and only if it contains a cycle.

For an example of how minors can arise naturally, note that if a graph is planar (meaning that it can be drawn in the plane in such a way that edges do not cross), then so is any minor of it. This is expressed by saying that the class of planar graphs is *minor closed*. Now, there is a theorem of Kuratowski that tells us which graphs are planar. One form that this theorem takes is the following statement: a graph is planar if and only if it does not have either K_5 or $K_{3,3}$ as a minor, where K_5 denotes the complete graph of order 5, and $K_{3,3}$ denotes the *complete bipartite* graph that consists of two sets of three vertices, with every vertex in one set joined to every vertex in the other set. Thus, the class of planar graphs is characterized by two *forbidden minors*.

Kuratowski's theorem tells us which graphs can be embedded into the plane. What happens for other surfaces? For example, it is easy to see that for any d the set of graphs that can be drawn on a d -holed torus is minor closed, but is there a finite set of forbidden minors in this case? To put it another way, is the set of obstructions to being embeddable into the d -holed torus only finite?

A special case of the Robertson-Seymour theorem states that the answer to this question is yes for any surface. But the theorem itself is much more general. It states that for *any* minor-closed class of graphs, there is a finite set of forbidden minors. In other words, for any minor-closed class \mathcal{G} there exist graphs G_1, \dots, G_k such that a graph G belongs to the family \mathcal{G} if and only if G does not have any G_i as a minor. There is also a pleasant form of the theorem (which is easily seen to be equivalent) that says that the class of all graphs is "well-quasi-ordered" by the minor relation: this means that given any sequence G_1, G_2, \dots of graphs there must exist one that is a minor of a later one.

It turns out that testing a graph for the presence of a given minor can be done reasonably fast, so that one amazing spin-off from the Robertson-Seymour theorem is that for any minor-closed class there is an efficient algorithm for checking whether or not a given graph belongs to the class. This has had a huge number of applications in routing problems and the like.

The actual proof of the Robertson-Seymour theorem is enormous: it was published in a sequence of twenty-two papers. Interestingly, it turns out that the case of graphs embeddable into a given surface plays a key role, as we now explain.

We will consider the form of the theorem mentioned above involving a sequence of graphs. So let us suppose for a contradiction that we have a "bad" sequence: that is, a sequence G_1, G_2, \dots for which no G_i is a minor of any later G_j . Let the number of vertices of the first graph G_1 be k . Since no later G_i has G_1 as a minor, it certainly follows that none of G_2, G_3, \dots has a complete minor of size k (or else we could delete some edges and obtain G_1). For this reason, Robertson and Seymour studied families of graphs that do not have a complete minor of size k . They were able to show that every graph that does not have a complete minor of size k may be built up in a certain way from graphs that are "nearly embeddable" into a fixed surface (that depends on the value of k). This means that in a certain sense that can be made precise the graph is not

too far from a graph that *is* embeddable into the surface. By some very deep arguments, they were able to show that the family of all such graphs (the graphs that can be built up from nearly embeddable graphs, for a given surface) has a finite number of forbidden minors, thereby proving the theorem.

V.36 The Three-Body Problem

The three-body problem can be simply stated: three point masses move in space under their mutual gravitational attraction; given their initial positions and velocities, determine their subsequent motion. Initially, it may come as a surprise that this is a difficult problem, since the analogous two-body problem can be solved fairly simply: more precisely, given any set of initial conditions, we can write down a formula, in terms of elementary functions (these are functions that can be built up using the basic operations of arithmetic, together with a few standard functions such as the EXPONENTIAL [III.25] and TRIGONOMETRIC [III.94] functions), that tells us the subsequent positions and velocities of the bodies. However, the three-body problem is a complicated nonlinear problem and it cannot be solved in this way, even if we are prepared to enlarge our stock of "standard functions" somewhat. NEWTON [VI.14] himself speculated that an exact solution "exceeds, if I am not mistaken, the force of any human mind," while HILBERT [VI.63], in his celebrated Paris address of 1900, put the problem in a category similar to FERMAT'S LAST THEOREM [V.12]. The problem can be extended to any number of bodies and in the general case it is known as the n -body problem.

Recall that the gravitational force of a particle P_1 on a particle P_2 has magnitude $k^2 m_1 m_2 / r^2$ (in suitable units), where k is the *Gaussian gravitational constant*, particle P_i has mass m_i , and the distance between the particles is r . The direction of this force on P_2 is toward P_1 (and there is a force of the same magnitude on P_1 in the direction of P_2). Recall also Newton's second law: force equals mass times acceleration. From these two laws we can easily derive the equations of motion for the three-body problem. Let the particles be P_1, P_2 , and P_3 . Write m_i for the mass of P_i , r_{ij} for the distance between P_i and P_j , and q_{ij} for the j th coordinate of the

position of P_i . Then the equations of motion are

$$\left. \begin{aligned} \frac{d^2 q_{1i}}{dt^2} &= k^2 m_2 \frac{q_{2i} - q_{1i}}{r_{12}^3} + k^2 m_3 \frac{q_{3i} - q_{1i}}{r_{13}^3}, \\ \frac{d^2 q_{2i}}{dt^2} &= k^2 m_1 \frac{q_{1i} - q_{2i}}{r_{12}^3} + k^2 m_3 \frac{q_{3i} - q_{2i}}{r_{23}^3}, \\ \frac{d^2 q_{3i}}{dt^2} &= k^2 m_1 \frac{q_{1i} - q_{3i}}{r_{13}^3} + k^2 m_2 \frac{q_{2i} - q_{3i}}{r_{23}^3}. \end{aligned} \right\} \quad (1)$$

Here, i runs from 1 to 3; thus, there are nine equations, all derived from the simple laws above. For instance, the left-hand side of the first equation is the component of the acceleration of P_1 in the i th direction, and the right-hand side is the component of the force acting on P_1 in this direction, divided by m_1 .

If the units are chosen so that $k^2 = 1$, then the potential energy V of the system is given by

$$V = -\frac{m_2 m_3}{r_{23}} - \frac{m_3 m_1}{r_{31}} - \frac{m_1 m_2}{r_{12}}.$$

Setting

$$p_{ij} = m_i \frac{dq_{ij}}{dt} \quad \text{and} \quad H = \sum_{i,j=1}^3 \frac{p_{ij}^2}{2m_i} + V,$$

we can rewrite the equations in the HAMILTONIAN FORM [IV.16 §2.1.3]

$$\frac{dq_{ij}}{dt} = \frac{\partial H}{\partial p_{ij}}, \quad \frac{dp_{ij}}{dt} = -\frac{\partial H}{\partial q_{ij}}, \quad (2)$$

which is a set of eighteen first-order differential equations. Since this set is easier to use, it is now generally preferred to (1).

A standard way of decreasing the complexity of a system of differential equations is to find an *algebraic integral* for it: that is, a quantity that will remain constant for any given solution and that can be expressed as an integral that gives rise to an algebraic dependence between the variables. This allows us to reduce the number of variables by expressing some of them in terms of others. The three-body problem has ten independent algebraic integrals: six of them tell us about the motion of the center of mass (three for the position variables and three for the momentum variables), three integrals express the conservation of angular momentum, and one expresses conservation of energy. These ten independent integrals were known to EULER [VI.19] and LAGRANGE [VI.22] in the middle of the eighteenth century, and in 1887 Heinrich Bruns, professor of astronomy at Leipzig, proved that there are no others, a result sharpened by POINCARÉ [VI.61] two years later. By the use of these ten integrals, together with the “elimination of the time” and the “elimination of the nodes” (a procedure first made explicit by JACOBI

[VI.35]), the original system of order eighteen can be reduced to one of order six, but it can be reduced no further. Hence, any general solution of (2) cannot be given by a simple formula: the best we can hope for is a solution in the form of an infinite series. It is not difficult to find series that work well enough for a limited time span: the problem is to find series that work for any initial configuration and for any time span, no matter how long. There is also the question of collisions. A complete solution to the problem has to take account of all possible motions of the bodies, including determining which initial conditions lead to binary and triple collisions. Since collisions are described by singularities in the differential equations, this means that to find a complete solution the singularities have to be understood.

This turns out to be a more interesting problem than one might think. It is obvious from the equations that a collision gives rise to a singularity, but it is less clear whether there can be any other kind of singular behavior. In the case of the three-body problem, the answer was supplied by Painlevé in 1897: the collisions are the only singularities. However, for more than three bodies the answer turned out to be different. In 1908 a Swedish astronomer, Hugo von Zeipel, showed that *noncollision singularities* can occur only if the system of particles becomes unbounded in a finite amount of time. A good example of such a singularity was found by Zhihong Xia for the five-body problem in 1992. In this case there are two pairs of bodies, the bodies in each pair having equal mass, and a fifth body with very small mass. The bodies in a pair move in very eccentric orbits parallel to the xy -plane, with the two pairs on opposite sides of this plane and rotating in opposite directions. A fifth particle is then added to the system. Its motion is confined to the z -axis and oscillates between the two pairs. Xia showed that the motion of the fifth particle forced the two pairs to move away from the xy -plane, but that it also came closer and closer to colliding with the pairs, giving it larger and larger bursts of acceleration, and that as this happened the two pairs were forced out to infinity in finite time.

As well as trying to solve the problem in general, one can look for interesting particular solutions. A *central configuration* is defined to be a solution in which the geometric configuration remains constant. The first examples were discovered by Euler in 1767: they were solutions in which the bodies always lie on a straight line and revolve with uniform angular velocity in circles or ellipses about their common center of mass. In

1772 Lagrange discovered solutions in which the bodies are always at the vertices of an equilateral triangle that rotates uniformly about the center of mass. For almost all sets of initial conditions for these solutions, the size of the triangle changes as it rotates so that each body describes an ellipse.

However, despite the discovery of the particular solutions and a century of unrelenting work on the problem, the mathematicians of the nineteenth century were unable to find a general solution. Indeed, the problem was considered so hard that in 1890 Poincaré was led to declare that he thought it impossible without the discovery of some significant new mathematics. But, contrary to Poincaré's expectation, less than twenty years later a young Finnish mathematical astronomer, Karl Sundman, using only existing mathematical techniques, astonished the mathematical world by obtaining uniformly convergent infinite series that mathematically "solved" the problem. Sundman's series, which are in powers of $t^{1/3}$, are convergent for all real t , except for the negligible set of initial conditions for which the angular momentum is zero. To deal with binary collisions, Sundman used the technique of *regularization*, or analytically extending a solution beyond the collision, but he was unable to deal with triple collisions because in order for such a collision to occur the angular momentum must be zero.

Although it was a remarkable mathematical achievement, Sundman's solution leaves many questions unanswered. It provides no qualitative information about the behavior of the system and, worse, because the series converges so slowly it is of no practical use. To determine the motion of the bodies for any reasonable period of time would require the summation of something of the order of $10^{8000000}$ terms, a calculation that is patently unrealistic. Thus, Sundman left plenty still to do, and work on the problem (and the related n -body problem) has continued up to the present day, with exciting results continuing to appear. One recent example is a convergent power-series solution for the general n -body problem, which was discovered by Don Wang in 1991.

Since the three-body problem itself proved so intractable, simplified versions were developed, of which the most famous is the one now known as the *restricted* three-body problem (the name is due to Poincaré), which was first investigated by Euler. In this case, two of the bodies (the *primaries*) revolve around their joint center of mass in circular orbits under the influence of their mutual gravitational attraction, while the third

body (the *planetoid*), which is assumed to have such small mass that the force it exerts on the other two bodies can be neglected, moves in the plane defined by the primaries. The advantage of this formulation is that the motion of the primaries can be treated as a two-body problem and is hence known; it remains only to investigate the motion of the planetoid, which can be done using perturbation theory. Although the restricted formulation might appear artificial, it provides a good approximation to real physical situations, such as, for example, the problem of determining the motion of the Moon around Earth given the presence of the Sun. Poincaré wrote extensively on the restricted problem, and the techniques he developed to tackle it led to his discovery of mathematical chaos, as well as laying the foundations for modern DYNAMICAL SYSTEMS [IV.14] theory.

PUP: can't add 'to' after this word as this means 'and also laid'.

Apart from its intrinsic appeal as a simple-to-state problem, the three-body problem has a further attribute that has contributed to its attraction for potential solvers: its intimate link with the fundamental question of the stability of the solar system. That is the question of whether the planetary system will always keep the same form as it has now, or whether, eventually, one of the planets will escape or, perhaps worse, experience a collision. Since bodies in the solar system are approximately spherical and their dimensions extremely small when compared with the distances between them, they can be considered as point masses. Ignoring all other forces, such as solar winds or relativistic effects, and taking only gravitational forces into account, the solar system can be modeled as a ten-body problem with one large mass and nine small ones, and it can be investigated accordingly.

Over the years, attempts to find a solution to the three-body problem (and the related n -body problem), have spawned a wealth of research. As a result, the importance of the problem is as much in the mathematical advances it has generated as in the problem itself. A notable example of this is the development of *KAM theory*, which provides methods for integrating perturbed Hamiltonian systems and obtaining results that are valid for infinite periods of time. This was developed in the 1950s and 1960s by KOLMOGOROV [VI.88], Arnold, and Moser.

Thurston's Geometrization Conjecture

See THE POINCARÉ CONJECTURE [V.28]

V.37 The Uniformization Theorem

The uniformization theorem is a remarkable classification of RIEMANN SURFACES [III.81]. Two surfaces are *biholomorphically equivalent* if there is a HOLOMORPHIC FUNCTION [I.3 §5.6] from one to the other that has a holomorphic inverse. If a Riemann surface is SIMPLY CONNECTED [III.95], then the uniformization theorem states that it is biholomorphically equivalent to the sphere, the Euclidean plane, or the HYPERBOLIC PLANE [I.3 §6.6]. These three spaces can all be viewed as Riemann surfaces, and they are all particularly symmetric: they have constant CURVATURE [III.80] (positive, zero, and negative, respectively); more generally, given any two points x and y in such a space, one can find a symmetry of the space that takes x to y , and one can ensure that a little arrow at x ends up pointing in any desired direction at y . Loosely speaking, these spaces “look the same from every point.”

It can be shown that an open subset of \mathbb{C} that is not the whole of \mathbb{C} cannot be biholomorphically equivalent to the sphere or to \mathbb{C} . Therefore, by the uniformization theorem, a simply connected open subset of \mathbb{C} that is not the whole of \mathbb{C} must be biholomorphically equivalent to the hyperbolic plane. This proves that any such set, no matter how irregular its boundary might be, can be mapped biholomorphically to any other. This result is called the *Riemann mapping theorem*. Biholomorphic maps are *conformal*: that is, if two curves in one set meet at an angle θ , then the angle between their images in the other set is also θ . So the Riemann mapping theorem implies that the interior of any simple closed curve can be mapped in an angle-preserving way to the open unit disk. Recall that one of the main models of the hyperbolic plane is Poincaré’s disk model. Thus, the hyperbolic metric on the disk together with the biholomorphic map that is given by the uniformization theorem can be used to define a hyperbolic metric on any simply connected proper open subset of \mathbb{C} .

If a Riemann surface is not simply connected, it is at least a QUOTIENT [I.3 §3.3] of a simply connected surface, namely its UNIVERSAL COVER [III.95]. For example, a torus is a quotient of the complex plane (in many possible ways that are topologically but not biholomorphically equivalent). Thus, the uniformization theorem tells us that a general Riemann surface is a quotient of the sphere, the Euclidean plane, or the hyperbolic plane. For a more detailed discussion of what such a quotient might be like, see FUCHSIAN GROUPS [III.28].

Waring’s Problem

See PROBLEMS AND RESULTS IN ADDITIVE NUMBER THEORY [V.30]

V.38 The Weil Conjectures

Brian Osserman

The Weil conjectures constitute one of the central landmarks of twentieth-century ALGEBRAIC GEOMETRY [IV.4]: not only was their proof a dramatic triumph, but they were the driving force behind a striking number of fundamental advances in the field. The conjectures treat a very elementary problem: how to count the number of solutions to systems of polynomial equations over finite FIELDS [I.3 §2.2]. While one might ultimately be more interested in solutions over, say, the field of rational numbers, the problem is far more tractable over finite fields, and LOCAL-GLOBAL PRINCIPLES [III.53] such as THE BIRCH-SWINNERTON-DYER CONJECTURE [V.4] establish strong, albeit subtle, relationships between the two cases.

Moreover, there are some basic questions that have nonobvious connections to the Weil conjectures. The most famous of these is the *Ramanujan conjecture*, which concerns the coefficients of $\Delta(q)$, one of the most fundamental examples of a MODULAR FORM [III.61]. We obtain the function $\tau(n)$ from the formula for $\Delta(q)$ as follows:

$$\Delta(q) = q \prod_{n=1}^{\infty} (1 - q^n)^{24} = \sum_{n=1}^{\infty} \tau(n) q^n.$$

RAMANUJAN [VI.82] conjectured that $|\tau(p)| \leq 2p^{11/2}$ for any prime number p . This is closely related to a statement on the number of ways of writing p as a sum of twenty-four squares. Work of Eichler, Shimura, Kuga, Ihara, and Deligne showed that in fact Ramanujan’s conjecture is a consequence of the Weil conjectures, so that Deligne’s proof of the latter in 1974 also resolved the former.

We begin with a brief historical summary of developments prior to WEIL [VI.93] and follow this with a more precise description of the statement of his conjectures. Finally, we sketch the ideas behind their proof.

1 An Auspicious Prologue

Our story begins with the seminal work of RIEMANN [VI.49] on the classical ZETA FUNCTION [IV.2 §3], which

we recall is defined by the sum

$$\zeta(s) = \sum_n \frac{1}{n^s}.$$

EULER [VI.19] had studied this function for real values of s , but RIEMANN, in his remarkable eight-page paper of 1859, went much further. He looked at complex values as well, and therefore had at his disposal the considerable resources of complex analysis. In particular, although the above sum for $\zeta(s)$ converges only for complex numbers s that have real part $\operatorname{Re}(s)$ strictly greater than 1, Riemann showed that the function itself can be extended to an analytic function defined on the entire complex plane, except at the point $s = 1$, at which it tends to infinity. He showed, moreover, that $\zeta(s)$ satisfies a certain functional equation relating $\zeta(s)$ to $\zeta(1-s)$, which introduced an important kind of symmetry around the line $\operatorname{Re}(s) = \frac{1}{2}$. Most famously (or infamously), he conjectured what is now known as THE RIEMANN HYPOTHESIS [I.4 §3]: that, aside from easily analyzed “trivial zeros” on the negative real axis, every zero of $\zeta(s)$ occurs on the line $\operatorname{Re}(s) = \frac{1}{2}$. Riemann’s motivation for studying $\zeta(s)$ was to analyze the distribution of prime numbers, but it fell to later authors (HADAMARD [VI.65], DE LA VALLÉE POUSSIN [VI.67], and Van Koch) to bring this vision to fruition. They used the zeta function to prove the PRIME NUMBER THEOREM [I.4 §3], which determined the asymptotic distribution of prime numbers, and also showed that the Riemann hypothesis is equivalent to a particularly strong upper bound for the error term in the prime number theorem.

At first glance, the Riemann hypothesis might appear to be completely special, a one-of-a-kind conjecture. However, it was not long before DEDEKIND [VI.50] generalized the Riemann hypothesis to a whole family of zeta functions, and in doing so opened the door to further generalization. Just as we can think of the complex numbers as being obtained from the real numbers by including a square root of -1 , that is, a root of the polynomial $x^2 + 1$, one can obtain a NUMBER FIELD [III.65], the fundamental object of study in ALGEBRAIC NUMBER THEORY [IV.1], from the field \mathbb{Q} of rational numbers by including roots of more general polynomials. For each number field K we have the ring of integers \mathcal{O}_K , which enjoys many of the same properties as the classical integers \mathbb{Z} . Starting from this observation, Dedekind defined a more general class of zeta functions, one for each such ring, which now bear his name. The classical zeta function $\zeta(s)$ was the Dedekind zeta function in the case $\mathcal{O}_K = \mathbb{Z}$. However, it was not at all straightforward to establish the existence of a functional equation

for Dedekind zeta functions: this was an open problem until 1917, when it was settled by Hecke, who showed at the same time that Dedekind zeta functions could be extended to the complex plane, thereby ensuring that the Riemann hypothesis makes sense for them as well.

With such ideas in the air, it was not long before geometry entered the picture. ARTIN [VI.86] first introduced zeta functions and the Riemann hypothesis for certain curves over finite fields in his 1923 thesis, noting that the ring of polynomial functions on such a curve shares precisely the properties of rings of integers that Dedekind used to define his zeta functions. Artin quickly observed first that his new zeta functions were strongly analogous to Dedekind zeta functions, and second that they were often more tractable: evidence for both observations is provided by the fact that he was able to check explicitly that the Riemann hypothesis was satisfied for a number of specific curves. The difference between the two situations is encapsulated as follows: while in the number field case one can think of the zeta function as counting primes, in the case of a function field the zeta function may be expressed in terms of the more geometric data of counting points on the given curve. In a 1931 paper F. K. Schmidt generalized Artin’s work, and exploited this geometry to prove a strong form of the functional equation for such zeta functions. And then, in 1933, Hasse proved the Riemann hypothesis in the special case of ELLIPTIC CURVES [III.21] over finite fields.

PUP: Tim says this sentence (and the use of singular) is fine. OK?

2 Zeta Functions of Curves

We now discuss in more detail the definition and properties of zeta functions associated with curves over finite fields, as well as the theorems of Schmidt and Hasse. Let \mathbb{F}_q denote the finite field with q elements, where $q = p^r$ for some prime number p and some positive integer r . The simplest case is when $q = p$, and \mathbb{F}_p is just the field of integers modulo p (see MODULAR ARITHMETIC [III.60]). More generally, we can obtain \mathbb{F}_q by adding roots of polynomials to \mathbb{F}_p , just as we do to \mathbb{Q} to obtain number fields; in fact, a single root of a single irreducible polynomial of degree r will do.

Artin studied a certain class of curves in the plane. Here, “plane” means \mathbb{F}_q^2 , that is, the set of all pairs (x, y) with x and y in \mathbb{F}_q . A curve C is simply the subset of these points where some polynomial $f(x, y)$ with coefficients in \mathbb{F}_q vanishes. Of course, if F is any field that contains \mathbb{F}_q , then the coefficients are also in F , so it makes sense to talk about $C(F)$, the curve in the larger

“plane” F^2 defined by the same equation $f(x, y) = 0$. If F is also a finite field, then $C(F)$ is obviously also finite. The finite fields F containing \mathbb{F}_q turn out to be the fields \mathbb{F}_{q^m} for $m \geq 1$. For each $m \geq 1$ let us define $N_m(C)$ to be the number of points belonging to the curve $C(\mathbb{F}_{q^m})$. The sequence $N_1(C), N_2(C), N_3(C), \dots$ is what we shall try to understand.

Given our plane curve C , we can define the *ring of polynomial functions* \mathcal{O}_C of C . This is simply the ring of polynomial functions on the plane (i.e., in two variables), modulo the EQUIVALENCE RELATION [I.2 §2.3] that two functions taking the same values on C should be considered the same. Formally, \mathcal{O}_C is simply the QUOTIENT [I.3 §3.3] ring $\mathbb{F}_q[x, y]/(f(x, y))$. Artin’s basic observation was that the definition of the Dedekind zeta function could be applied equally well to the ring \mathcal{O}_C , yielding a zeta function $Z_C(t)$ associated with C . However, in our geometric context we have the following equivalent and more elementary formula, which explicitly relates $Z_C(t)$ to the number of points over finite fields:

$$Z_C(t) = \exp \left(\sum_{m=1}^{\infty} N_m(C) \frac{t^m}{m} \right). \quad (1)$$

Schmidt generalized Artin’s definition to all curves over finite fields, and gave an elegant description of the zeta function for curves, bearing out Artin’s observations in the cases he was able to compute. The nicest form of Schmidt’s theorem applies to curves that satisfy two additional conditions. The first condition is that, rather than considering the curve C in the plane, we will want to “compactify” it by considering instead a *projective* curve; we can think of this as adding some “points at infinity,” thus increasing $N_m(C)$ slightly. Second, we will want to impose a technical condition of *smoothness* on C , which is analogous to asking that C be a MANIFOLD [I.3 §6.9].

In order to state Schmidt’s result, recall that there is a notion of the *genus* g (see ALGEBRAIC GEOMETRY [IV.4 §10]) of a smooth projective curve C , which can be defined to be the dimension of the space of differentials on C , or, if C is a complex curve, as the “number of holes” in the space obtained from the analytic topology on C . By extending certain classical results in algebraic geometry to more general fields, Schmidt proved that, for a smooth projective curve C over \mathbb{F}_q of genus g , we have

$$Z_C(t) = \frac{P(t)}{(1-t)(1-qt)}, \quad (2)$$

where $P(t)$ is a polynomial of degree $2g$ with integer coefficients. Furthermore, he proved a functional equa-

tion in terms of the substitution $t \mapsto 1/qt$. If we set $t = q^{-s}$, this gives a functional equation for the substitution $s \mapsto 1-s$, as in Riemann’s original work. The Riemann hypothesis for C is then the statement that the roots of $Z_C(q^{-s})$ all have $\operatorname{Re}(s) = \frac{1}{2}$, or, equivalently, that the roots of $P(t)$ all have norm equal to $q^{-1/2}$. It is an elementary observation that this is equivalent to the assertion that $|N_m(C) - q^m + 1| \leq 2g\sqrt{q^m}$, for all $m \geq 1$.

The next step in exploiting the geometric nature of zeta functions of curves is the observation that if F is a finite field containing \mathbb{F}_{q^m} , then the points with coordinates in \mathbb{F}_{q^m} are the fixed points of a function called the *Frobenius map*, which is the map Φ_{q^m} that sends a point $(x, y) \in F^2$ to the point (x^{q^m}, y^{q^m}) . It is a simple extension of FERMAT’S LITTLE THEOREM [III.60] that if $t \in \mathbb{F}_{q^m}$, then $t^{q^m} = t$. Moreover, the converse holds: if F is a field containing \mathbb{F}_{q^m} , and $t \in F$ satisfies $t^{q^m} = t$, then $t \in \mathbb{F}_{q^m}$. This follows because in any field, and in particular in F , the polynomial $t^{q^m} - t$ can have at most q^m roots, which must then be precisely the elements of \mathbb{F}_{q^m} . It immediately follows that a point $(x, y) \in F^2$ is a fixed point of Φ_{q^m} if and only if $(x, y) \in \mathbb{F}_{q^m}^2$. Moreover, it is elementary that $(s+t)^{q^m} = s^{q^m} + t^{q^m}$, if s, t are in any field containing \mathbb{F}_p . Because the coefficients of $f(x, y)$ are in \mathbb{F}_{q^m} , it follows that if $f(x, y) = 0$, then

$$f(\Phi_{q^m}(x, y)) = f(x^{q^m}, y^{q^m}) = (f(x, y))^{q^m} = 0,$$

so we see that Φ_{q^m} gives a map from C to itself. Thus, one might hope to study $C(\mathbb{F}_{q^m})$ by analyzing more generally what one can say about the fixed points of maps from C to itself. Hasse successfully applied this point of view to prove the Riemann hypothesis in the case $g = 1$, which is to say the case of elliptic curves. Moreover, we will see that this perspective is woven throughout the fabric of the rest of our story, not only inspiring Weil to make his conjectures, but also suggesting the techniques that ultimately led to their proof.

3 Enter Weil

In 1940 and 1941, Weil gave two proofs of the Riemann hypothesis for curves over finite fields. Or, to be more accurate, he described two proofs: they both relied on fundamental facts in algebraic geometry which had been proved by analytic methods for varieties over the complex numbers, but which had not been proved rigorously in the case of arbitrary base fields. It was largely in order to address this deficiency that Weil wrote his *Foundations of Algebraic Geometry*, which appeared in

1948 and allowed both of his earlier proofs to be made rigorous.

Weil's book constituted a watershed in algebraic geometry, as it introduced for the first time the notion of an *abstract* algebraic variety. Previously, a variety was always a global object, in that it was defined by a single collection of polynomial equations, in either affine or projective space. Weil realized that it would be helpful to have a corresponding locally defined concept, so he introduced abstract algebraic varieties, which are obtained by gluing together affine algebraic varieties in much the same way that manifolds in topology are obtained by gluing together open subsets of affine space. The notion of an abstract variety played a fundamental role in formalizing Weil's proofs, and was also an important precursor to Grothendieck's immensely successful theory of SCHEMES [IV.5 §3].

The following year, in a remarkable paper in the *Bulletin of the American Mathematical Society*, Weil went further, studying zeta functions $Z_V(t)$ associated with higher-dimensional varieties V over finite fields, and taking as his definition the formula (1). While the situation is more complicated in this context, the behavior conjectured by Weil was nonetheless strikingly similar, an utterly natural extension of the case of curves:

- (i) $Z_V(t)$ is a rational function of t ;
- (ii) more explicitly, if $n = \dim V$, we can write

$$Z_V(t) = \frac{P_1(t)P_3(t) \cdots P_{2n-1}(t)}{P_0(t)P_2(t) \cdots P_{2n}(t)},$$

where each root of each $P_i(t)$ is a complex number of norm $q^{-i/2}$;

- (iii) the roots of $P_i(t)$ are interchanged with the roots of $P_{2n-i}(t)$ under the substitution $t \mapsto 1/q^n t$;
- (iv) if V is the reduction modulo p of a variety \tilde{V} defined over a subfield of \mathbb{C} , then $b_i = \deg P_i(t)$ is the i th Betti number of \tilde{V} using the usual topology.

The last part of (ii) is known as the Riemann hypothesis, while (iii) constitutes a functional equation for the substitution $t \mapsto 1/q^n t$. Betti numbers are a well-known invariant from ALGEBRAIC TOPOLOGY [IV.6]: if we return to Schmidt's theorem (2) in the case of curves, the degrees $1, 2g, 1$ of $1-t, P(t), 1-qt$ are precisely the Betti numbers of a complex curve of genus g .

4 The Proof

Weil's conjectures were inspired by a very intuitive topological picture, derived from considering $V(\mathbb{F}_{q^m})$ as the set of fixed points of Φ_{q^m} . Forgetting for the

moment that Φ_{q^m} makes sense only over finite fields, if we imagine that V were defined over the complex numbers, then by using the complex topology we could study the fixed points of Φ_{q^m} by the LEFSCHETZ FIXED-POINT THEOREM [V.13 §3], obtaining a formula in terms of the action of Φ_{q^m} on the COHOMOLOGY GROUPS [IV.6 §4]. Indeed, we could deduce the factorization in (ii) almost immediately (and in particular the rationality asserted in (i)), with each factor $P_i(t)$ corresponding to the action of Frobenius on the i th cohomology group, and we would also have $\deg P_i(t)$ given by the i th Betti number of V . Moreover, the functional equation would follow from a concept known as POINCARÉ DUALITY [III.19 §7].

It was not long before it became clear that such cohomological arguments might become more than just motivation: there could be a cohomology theory for algebraic varieties over finite fields that would mimic the properties of the classical topological theory and would allow one to prove the Weil conjectures. Such a cohomology theory is now known as a *Weil cohomology*. Serre was the first to seriously attempt to develop such a theory, but he had only limited success. In 1960, Dwork provided a brief detour by using p -ADIC ANALYSIS [III.53] to prove parts (i) and (iii) of the conjectures: that is, the rationality and the functional equation. Shortly thereafter, building on comments of Serre and in collaboration with Artin, Grothendieck proposed and developed a candidate for a Weil cohomology, the *étale cohomology*. Indeed, he noted that one could in fact extend the list of desired properties of a Weil cohomology in such a way that the Weil conjectures would follow almost immediately. These properties were known but extremely difficult in the classical case, and included the "hard Lefschetz theorem." In a burst of optimism, Grothendieck referred to them as the "standard conjectures," and envisioned that the Weil conjectures would ultimately be proved through them.

However, the final chapter of the story did not go entirely according to Grothendieck's plan. His student Deligne set about working on the problem, and was ultimately able to complete an exceedingly subtle and intricate proof using induction on the dimension of the variety. The étale cohomology played an absolutely fundamental role in Deligne's proof, but he also introduced other ideas into the picture, most notably a classical geometric construction of Lefschetz, as well as some work of Rankin on the Ramanujan conjecture. In the end, he was able to conclude the hard Lefschetz

theorem from his work, but the rest of the standard conjectures remain unsolved to this day.

Acknowledgments. I would like to thank Kiran Kedlaya, Nicholas Katz, and Jean-Pierre Serre for their helpful correspondence.

Further Reading

- Dieudonné, J. 1975. The Weil conjectures. *Mathematical Intelligencer* 10:7–21.
- Katz, N. 1976. An overview of Deligne's proof of the Riemann hypothesis for varieties over finite fields. In *Mathematical Developments Arising from Hilbert Problems*, edited by F. E. Browder, pp. 275–305. Providence, RI: American Mathematical Society.
- Weil, A. 1949. Numbers of solutions of equations in finite fields. *Bulletin of the American Mathematical Society* 55: 497–508.

Part VI

Mathematicians

VI.1 Pythagoras

b. Samos, Ionia (now Samos, Greece)?, ca. 569 B.C.E.;
d. Metapontum, Magna Graecia (now Metaponto, Italy)?, ca. 494 B.C.E.
Incommensurability; theorem of Pythagoras

One of the most elusive figures of antiquity, Pythagoras is famous not just for his alleged mathematical achievements: it has been claimed that he had a golden thigh and that he issued a prescription against broad beans. Few things about him can be taken as historical facts, but we can be reasonably confident that he lived in around the sixth century B.C.E. in Greek southern Italy and that he established a group of followers, the Pythagoreans, who shared not just beliefs, but also dietary habits and a code of behavior. The existence of anecdotes about splinter Pythagoreans who revealed secrets to outsiders and were accordingly punished suggests that they were far from constituting a completely homogeneous group.

After a peak period in the late fifth century B.C.E., the Pythagoreans dispersed, probably as a result of their involvement in the public life of various city-states. The impact of their theories about the universe and the soul was very long-lived, though, and can be felt in Plato, Aristotle, and later authors. From the third century B.C.E. until well into late antiquity, a stream of texts was produced that purported to be by Pythagoras or his immediate successors. Indeed, historians talk of a neo-Pythagorean philosophical movement, sometimes associated with neo-Platonism.

The name of Pythagoras and his school is most commonly linked to the theorem establishing that the square on the hypotenuse of a right-angled triangle is equivalent to the sum of the squares on the other two sides. In fact, there is some evidence that the mathematical property expressed by the theorem was known in Mesopotamia long before Pythagoras's time; the ancient sources attributing the result to him are late and not entirely reliable, and no actual proof of

the theorem is found before Euclid's *Elements*. While the proof itself may predate EUCLID [VI.2], there is no solid reason to connect it to Pythagoras.

Similarly, the discovery of the incommensurability of the side and the diagonal of a square, often attributed to Pythagoreans, may have been made earlier in Mesopotamia, and the earliest full proof in a Greek context belongs to a later period.

Pythagoras's real contribution to mathematics lies elsewhere. The Pythagoreans are credited by Aristotle with the theory that "things themselves are numbers." One interpretation is that they believed that mathematics offered a key to understanding reality, whether this reality was conceived to have an underlying geometrical structure (as in Plato's *Timaeus*), or whether it was simply seen as ordered and "in proportion." Indeed, Pythagoreans are plausibly credited with a strong interest in formulating the numerical ratios of musical concords and harmony. They connected the harmonious sound produced by, say, the plucking of a string with the fact that the musician plucked it at specific, mathematically expressible points. Breaking the mathematical proportion between the points on the string unsettled the sound produced. The heavenly bodies themselves, according to the Pythagoreans, produced music, thanks to their mathematical, and therefore orderly, arrangement. Understand the mathematics, and you will grasp the structure of reality: this insight is perhaps Pythagoras's true legacy.

Further Reading

- Burkert, W. 1972. *Lore and Science in Ancient Pythagoreanism*. Cambridge, MA: Harvard University Press. (Revised English translation of 1962 *Weisheit und Wissenschaft: Studien zu Pythagoras, Philolaos und Platon*. Nürnberg: H. Carl.)
- Zhmud, L. 1997. *Wissenschaft, Philosophie und Religion im frühen Pythagoreismus*. Berlin: Akademie.

Serafina Cuomo

PUP: question marks added to some of the birth and death places for the first four mathematicians in the section to indicate that the places are uncertain. OK? Or do you have a suggestion for a better way of indicating this? Also, 'c.' changed to 'ca.' to match style elsewhere in Companion. OK?

VI.2 Euclid

b. Alexandria, Egypt?, ca. 325 B.C.E.; d. Alexandria?, ca. 265 B.C.E.
Deduction; postulate; reductio ad absurdum

Nothing is known about Euclid's life. In fact, his major work, the *Elements*, is now seen as a rather loose collection, with no strong authorial voice and no clear way of determining what, if any, Euclid's original contributions were. Born in the cultural climate of Ptolemaic Alexandria, the text probably aimed at systematizing the current knowledge in some mathematical areas.

The *Elements* covers plane geometry (including the squaring of any rectilinear figure, the bisection of an arc, the inscription and circumscription of polygons in circles, the finding of a mean proportional), solid geometry (e.g., the ratio of spheres to one another, the five regular solids), and arithmetic, from relatively simple (the properties of odd and even numbers, prime number theory) to more complex (commensurable and incommensurable lines, binomials and apotomes).

The title hints at the foundational character of the text, which starts with definitions of mathematical objects (e.g., point, straight line, scalene triangle), postulates (e.g., all right angles are equal to one another) and common notions (e.g., the whole is greater than the part). These initial premises are not demonstrated—whether some postulates are demonstrable spawned debate in antiquity, and later led to non-Euclidean geometries. In a style which has been termed axiomatic-deductive, proofs tend to be general rather than specific; they use a restricted set of formulaic expressions, refer to a lettered diagram, and each of their steps is justified by appeal to undemonstrated premises, to previous proofs, or to very simple notions, such as the principle of the excluded middle. Some proofs use *reductio ad absurdum*: instead of directly showing that something is the case, they proceed to show that any alternative is impossible.

There are parts of the book that reveal the presence of different, less abstract, demonstrative procedures. For instance, one of the theorems establishing criteria for two triangles to have the same area refers to one triangle being “superimposed” on the other, with the reader effectively invited to verify that their areas are indeed equal. The appeal is to a mental operation, which is quite different from the logical step-by-step method found elsewhere. Again, book IX contains propositions on odd and even numbers, which are often seen as vestiges of Pythagorean mathematics, to be

demonstrated with the help of pebbles. The coexistence itself of arithmetic and geometry has been puzzling for some historians, who have proposed a notion of “geometric algebra,” so that book II, ostensibly about squares and rectangles built on segments of straight lines, would in fact foreshadow modern equations.

As well as works on astronomy, optics, and music, the *Data*, which is about solving geometrical problems on the basis of some elements that are already given, is also attributed to Euclid. His fame is, however, inextricably linked to the *Elements*. The very absence of a strong authorial voice has perhaps facilitated other mathematicians' interaction with the text, which has been appropriated, added to, interfered with, and commented upon since antiquity. This very plasticity helped to make it possibly the most popular mathematical book of all time. (For more about its impact on the early development of mathematics, see GEOMETRY [II.2], THE DEVELOPMENT OF ABSTRACT ALGEBRA [II.3], and THE DEVELOPMENT OF THE IDEA OF PROOF [II.6].)

Further Reading

- Euclid. 1990–2001. *Les Éléments d'Euclide d'Alexandrie; Traduits du Texte de Heiberg*, general introduction by M. Caveing, translation and commentary by B. Vitrac, four volumes. Paris: Presses Universitaires de France.
- Netz, R. 1999. *The Shaping of Deduction in Greek Mathematics. A Study in Cognitive History*. Cambridge: Cambridge University Press.

Serafina Cuomo

VI.3 Archimedes

b. Siracusa, Magna Graecia (now Syracuse, Italy), ca. 287 B.C.E.;
 d. Siracusa, 212 B.C.E.
Area of the circle; centers of gravity; method of exhaustion; volume of the sphere

Archimedes' life was as spectacular as his scientific achievements: various sources attest that he built a ship, a cosmological model, and magnificent catapults with which he defended his native Syracuse during the Second Punic War. The Roman besiegers eventually took the city by deceit, and Archimedes was killed in the ensuing pillage. According to legend, his tomb was engraved with a sphere inscribed in a cylinder, to mark one of his most famous discoveries. Indeed, the first part of his *Sphere and Cylinder* reaches a climax with a proof that the volume of every sphere is two thirds of that of the cylinder circumscribing it. Archimedes'

interest in establishing the volume or area of curved figures is also attested by his discovery of the area of the circle and of a sphere, and by treatises on spiral curves, conoids, and paraboloids, and on the *Quadrature of the Parabola*.

While following an axiomatic-deductive framework, Archimedes' style is distinctive. Many of his theorems about curved figures use the so-called METHOD OF EXHAUSTION [II.6 §2].

Take the problem of determining the area of a circle. Archimedes accomplished this by showing that it had the same area as that of a certain right-angled triangle. Since it was known how to calculate the area of a triangle, he was "reducing" a problem whose solution was unknown to one whose solution was known. Rather than establishing this directly, he proves that the area of the circle can be neither larger than nor smaller than the area of the triangle, so that only one possibility remains: that they are equal. This is achieved, here and in general, by inscribing and circumscribing rectilinear figures to the curvilinear figure under investigation, thus getting closer and closer to it. The leap from closer and closer approximation to equivalence of a rectilinear and curvilinear figure, however, can be accomplished only indirectly, by excluding the other possibilities. Such arguments usually employ a lemma, already found in Euclid, to the effect that if we start with a quantity and replace it by a quantity at most half as large, and then repeat this, then what remains can be made as small as we please.

Archimedes' output also includes *The Sand-Reckoner*, about astronomy and arithmetic, and works on the centers of gravity of plane figures and on bodies immersed in a fluid.

Above all, Archimedes provides unique insights into the processes of ancient Greek mathematics. The second part of *Sphere and Cylinder* contains problems about constructing given solid bodies. Several of the proofs are in two parts: analysis and synthesis. In the analysis, the result one wants to establish is taken as proved, and consequences are drawn from it, until one hits upon a result that is already proved elsewhere, and the process is then reconstituted in reverse (the "synthesis"). The recently rediscovered *Method* (addressed to Eratosthenes) reveals that Archimedes arrived at some of his most famous results, e.g., the area of a segment of a parabola, by imagining the two objects involved (say, a segment of a parabola and a triangle) as divided up into an infinite number of slices and lines,

then placed at the two ends of a balance and set in equilibrium with each other. Archimedes underlined that this heuristic procedure was not a strict proof, but that only makes the *Method* all the more valuable a glimpse into the mind of a great mathematician.

Further Reading

Archimedes. 2004. *The Works of Archimedes: Translation and Commentary. Volume 1: The Two Books On the Sphere and the Cylinder*, edited and translated by R. Netz. Cambridge: Cambridge University Press.

Dijksterhuis, E. J. 1987. *Archimedes*, with a bibliographical essay by W. R. Knorr. Princeton, NJ: Princeton University Press.

Serafina Cuomo

VI.4 Apollonius

b. Perge, Pamphylia (now Perge, Turkey), ca. 262 B.C.E.;

d. Alexandria, Egypt?, ca. 190 B.C.E.

Conic sections; diorism; locus problems

The *Conics*, in eight books, only seven of which have come down to us, has had fewer modern readers than other recognized masterpieces of Greek mathematics: it is complex, difficult to summarize, and easy to mis-translate into modern algebraic notation. Apollonius of Perge also wrote about arithmetic and astronomy, but none of these works survive. The letters prefacing six of the surviving books indicate that he was a highly esteemed member of a network of mathematicians, to whom he sent his results. He refers to the fact that various versions of his *Conics* were circulated, the latest probably incorporating his correspondents' feedback. Knowledge of the parabola, hyperbola, and ellipse predates Apollonius (we find conics in Archimedes), but his is the first known systematic account of these curves, which were of interest both in themselves and because they could be used as auxiliary lines for the solution of problems such as the trisection of an angle or the duplication of the cube.

Apollonius himself declares that the first four books of the *Conics* are an introduction to the subject, and indeed he starts with definitions of the cone and its various parts. The parabola, hyperbola, and ellipse are not introduced until later, so that their origin (from a plane cutting a cone or a conic surface at different angles) is already accompanied by a statement of their properties, which are further and fully explored in the next three books. These include theorems on tangents,

asymptotes, and axes; constructions of conic sections on the basis of certain data; and an account of the conditions under which conics can intersect in the same plane.

The nonelementary books, which exist only in Arabic, contain treatments of maximum and minimum lines within the sections, the construction of conic sections equal or similar to a given conic section (including the theorem that all parabolas are similar), and “diorismic theorems.” These are propositions that set the limits of possibility of a construction, or the limits of validity of some property of a geometrical configuration, given a certain number of known positions or known objects at the outset. Indeed, several of the propositions in the *Conics* are about loci, i.e., geometrical configurations consisting of all the points sharing a certain family of properties. Apollonius criticizes EUCLID [VI.2] for not having provided an exhaustive solution to the construction of the three-line and four-line locus (configurations of three or four lines, arranged so that they have specific properties).

In terms of demonstrative methods, Apollonius is in the axiomatic-deductive mold: general enunciations, lettered diagrams, each step justified by appeal to undemonstrated premises or previous proofs. Instead of indirect methods, we find a real mastery of the intricacies (and power) of proportion theory. At the same time, his propositions easily lend themselves to the consideration of different subcases: when, for instance, a certain line falls inside, outside, or on the vertex of a conical surface. Apollonius, in other words, combines a systematic approach with an almost playful fascination with exploring the possibilities of mathematical objects and their properties under varying circumstances.

Further Reading

- Apollonius. 1990. *Conics*, books V–VII. *Arabic Translation of the Lost Greek Original in the Version of the Banu Musa*, edited with translation and commentary by G. J. Toomer, two volumes. New York: Springer.
- Fried, M. N., and S. Unguru. 2001. *Apollonius of Perga's Conica: Text, Context, Subtext*. Leiden: Brill.

Serafina Cuomo

VI.5 Abu Ja'far Muhammad ibn Mūsā al-Khwārizmī

b. Unknown, 800; d. Unknown, 847
Arithmetic; algebra

Al-Khwārizmī, or possibly his ancestors, came from Khwārizm (the modern region of Khorezm in Uzbekistan, also known as Khiva). Most of his life was spent as a scholar at the House of Wisdom, Baghdad, where he produced works on astronomy, mathematics, and geography. Of his mathematical works, two have come down to us, one on arithmetic and one on algebra.

The arithmetical work, which did not survive in Arabic and is known only through Latin translations, was the means by which Hindu numerals were transmitted to the West, as well as the corresponding methods of arithmetical calculation. Although the text was clearly based on Indian writings, in Europe the techniques became particularly associated with al-Khwārizmī's name in the form of *algorism* (from which the modern term “algorithm” is derived).

Al-Khwārizmī's *al-Kitāb al-mukhtaṣar fī ḥisāb al-jabr wa'l-muqābala* (“The compendious book on calculation by completion and balancing”) became the starting point for the subject of algebra for Islamic mathematicians. A work of elementary practical mathematics, it is written in three parts: one was devoted to solving equations, one to practical mensuration (areas and volumes), and one to problems that arose mainly from the complicated Islamic laws of inheritance (involving arithmetic and simple linear equations). No algebraic symbolism is employed: everything, including numerals is expressed in words. The text opens with a brief discussion of the place-value system and then deals with equations of the first and second degrees. Remarkably, al-Khwārizmī did not regard these equations just as a means for solving problems, as his predecessors had done, but studied them in their own right, classifying them into six separate types. In modern notation these are

$$\begin{aligned} ax^2 &= bx, & ax^2 &= b, & ax &= b, \\ ax^2 + bx &= c, & ax^2 + c &= bx, & ax^2 &= bx + c, \end{aligned}$$

where a , b , and c are positive integers. The different types are necessary because al-Khwārizmī did not recognize the existence of either negative numbers or zero as coefficients. Not only did al-Khwārizmī give proofs that his methods worked, which in itself was not standard at the time, but the proofs he gave were geometrical ones. That is, they were not classical Greek proofs but geometrical demonstrations of the validity of his methods.

The key word of the Arabic title, *al-jabr* (“completion” or “restoration”), which refers to restoring all the

terms to a standard form, eventually came into common usage in the West as *algebra*. It is, however, doubtful whether al-Khwārizmī's work was the first Islamic work bearing that name.

Further Reading

Berggren, J. L. 1986. *Episodes in the Mathematics of Medieval Islam*. New York: Springer.

VI.6 Leonardo of Pisa (known as Fibonacci)

b. Pisa, Italy, ca. 1170; d. Pisa, Italy, ca. 1250

Son of Pisan merchant; studied mathematics under Muslim teachers in North Africa and traveled throughout the Mediterranean meeting with Islamic scholars; awarded an annual stipend in 1240 by the city of Pisa in recognition of his teaching and other services

One of the earliest European writers on algebra, Fibonacci is most famous for his *Liber Abaci* ("Book of calculation"), which first appeared in 1202 and was largely responsible for the spread of the Hindu-Arabic numerals throughout Europe. The book contained not only rules for computing with the Hindu-Arabic numerals but also a large number of problems of various kinds, the best known of which was his "rabbit problem." This problem asks how many pairs of rabbits will be produced in a year, beginning with a single pair, if in every month each pair produces a new pair which becomes productive from the second month on. The number F_n of pairs there will be in the n th month is the number of pairs there were in the previous month plus the number of breeding pairs, and the latter is the number of rabbits there were in the previous month but one. This leads to the rule $F_n = F_{n-1} + F_{n-2}$. Starting with $F_0 = 0$ and $F_1 = 1$, we obtain the sequence 0, 1, 1, 2, 3, 5, 8, 13, ... of *Fibonacci numbers*. It can be shown that $\lim_{n \rightarrow \infty} F_{n+1}/F_n = \phi$, where $\phi = (1 + \sqrt{5})/2$ is the golden ratio.

VI.7 Girolamo Cardano

b. Pavia, Italy, 1501; d. Rome, 1576

Teacher of mathematics, Milan (1534–43); Professor of Medicine: Pavia (1543–60), Bologna (1562–70); imprisoned for heresy (1570–71)

Cardano's great treatise, the *Ars Magna* (1545), laid the foundations for European algebra and remained the most comprehensive and systematic work on algebra for more than a century after it was published. It contained many new ideas, including methods (not all Cardano's own) for solving cubic and quartic equations,

all written without mathematical notation. Cardano's own great insight was to recognize the existence of relations between the roots and the coefficients of an equation; in this he was unprecedented. He also showed a greater readiness than most of his contemporaries to contemplate the square roots of negative numbers. He is remembered today for "Cardano's rule" for solving cubic equations of the form $x^3 + cx = d$, where c and d are positive (he was unable to solve the *casus irreducibilis*, the case when c is negative).

VI.8 Rafael Bombelli

b. Bologna, Italy, 1526; d. Probably Rome, after 1572

Engineer-architect for the Roman nobleman Alessandro Rufini, later Bishop of Melfi

Bombelli was prompted to write his *Algebra* (1572) by a desire to make CARDANO'S [VI.7] *Ars Magna* (1545) accessible to the less sophisticated reader. The *Algebra*, which contains a systematic treatment of quadratics, cubics, and quartics, is noted for its advances in mathematical notation—it was the first printed text to include a notation for exponents—and for its role in disseminating awareness of the work of Diophantus. Above all, the *Algebra* was renowned for solving certain special cases of the so-called *casus irreducibilis* of the cubic, those in which Cardano's rule appears to give rise to a complex or "impossible" solution. Cardano was aware that what we today call complex numbers (numbers of the form $a + b\sqrt{-1}$) could arise in the solution of quadratic equations. Bombelli made the important discovery that what at first sight appears to be a complex root of a cubic equation may in fact be a real root, because the imaginary parts cancel each other out. The *Algebra* included the first extensive discussion of complex numbers and Bombelli formulated the four basic operations of arithmetic for them.

VI.9 François Viète

b. Fontenay-le-Comte, France, 1540; d. Paris, 1603
Trigonometry; algebraic analysis; classical problems; numerical solution of equations

Viète obtained a bachelor's degree in law in 1560 from the University of Poitiers, but left the profession from 1564 to 1568 to oversee the education of Catherine de Parthenay, daughter of a local aristocratic family. His earliest scientific writings were his lectures to Catherine. He spent the remainder of his life in high public office, apart from a period between 1584 and 1589

PUP: I'm sure we have discussed this before but I now wonder whether the information in italics in the micro entries should go above the hline as it does for the non-micro entries? I know the information is slightly different but maybe the differentiation isn't as important as that it looks odd (compare the four articles that start on this page for example)?

when he was banished from the court in Paris for political and religious reasons. He died in Paris in 1603. Throughout his life, it was only during the time he had free from official duties that he was able to devote himself to mathematics.

The work for which Viète is best known appeared during the 1590s, beginning with *In Artem Analyticem Isagoge* ("Introduction to the analytic art") in 1591. In the *Isagoge* Viète began to combine classical Greek geometry with algebraic methods that had originated from Islamic sources, and in doing so laid the foundations for the algebraic approach to geometry. Viète saw that the symbols in equations (traditionally variants of R , Q , and C , for the unknown, its square, and its cube) could represent either numbers or geometric quantities, and that this was potentially a powerful tool for analyzing and solving geometric problems.

Viète's understanding of analysis was based on his reading of the *Synagoge* ("Collection") of Pappus (early fourth century C.E.), where analysis was described as a method of investigating a problem by assuming that the solution is in some sense known, as we would do now by representing the solution by a symbol and carrying out mathematical manipulations involving that symbol. Algebra achieves this by regarding all quantities, known or unknown, as of equal status; equations are then formed from predated conditions (a process Viète called *zetetics*), and solved to produce the unknown quantity in terms of those given (*exegetics*). For Viète the final step in geometric problems was to provide a specific construction for the solution: this was the geometric *synthesis* arising from the preceding algebraic *analysis*.

In several further treatises, mostly written or published around 1593, Viète taught the necessary skills of forming equations and carrying out the corresponding geometric constructions, and these books together made up his *Opus Restitutae Mathematicae Analyseos, seu Algebra Nova* ("The work of restored mathematical analysis, or the new algebra"), which he offered with the famous and ambitious hope of leaving no mathematical problem unsolved (*nullum problema non solve*). For most of the seventeenth century, algebra continued to be known as the "analytic art," or simply "analysis."

Recognizing that not all equations could be solved algebraically, Viète also put forward a method of numerical solution based on successive approximations. This was the first appearance of such techniques in Europe, and was important not only for practical purposes, but also because it rapidly led to a deeper

understanding of the relationships between roots and coefficients of equations.

Viète's style of writing is wordy and often obscure, thanks in part to his liking for technical Greek terms. In his algebraic treatises, however, he devised some rudimentary notation. It had long been the case that rules for solving equations were presented through particular examples that were understood to represent a general class, but Viète took the step of replacing known quantities by consonants B, C, \dots , and unknowns by vowels A, E, \dots , so that numbers were replaced throughout by letters, or "species." However, he had no simple or systematic way of denoting powers (for squares and cubes he used the verbal *A quadratus* and *A cubus*), and his connectives ("added to," "equals," and so on) were also written in words, so that his algebra was still very far from symbolic.

One of the first people to study Viète's work in depth was Thomas Harriot in England, who, through study of Viète's numerical method shortly after 1600, discovered that polynomials could be written as products of linear and quadratic factors, a major breakthrough in the understanding of equations. Harriot also rewrote much of Viète's mathematics in what is essentially modern algebraic notation. In France, Viète's work was taken up in the 1620s by FERMAT [VI.12], who was profoundly influenced by it. DESCARTES [VI.11], on the other hand, denied that he had ever read either Viète or Harriot, though in the 1630s he developed a number of very similar ideas.

Viète and his immediate successors dealt only with equations of finite degree. Only much later in the seventeenth century with the work of NEWTON [VI.14] was analysis extended to include what were thought of as infinite equations, or what we would now call infinite series, hence bringing the word "analysis" much closer to its modern meaning.

Jacqueline Stedall

VI.10 Simon Stevin

b. Bruges, Belgium, 1548; d. The Hague, the Netherlands, 1620

Mathematics and science tutor to Maurice of Nassau, Prince of Orange

The Flemish mathematician and engineer Simon Stevin is remembered for his study of decimal fractions. Although he was not the first to use decimal fractions (they are found in the work of the tenth-century Islamic

mathematician al-Uqlidīsī), it was his tract *De Thiende* (“The tenth”), published in 1585 and translated into English (as *Disme: The Art of Tenths, or Decimall Arithmetike Teaching*) in 1608, that led to their widespread adoption in Europe. Stevin, however, did not use the notation we use today. He drew circles around the exponents of the powers of one tenth: thus he wrote 7.3486 as 7⑩3⑩4⑩28⑩36⑩4. In *De Thiende* Stevin not only demonstrated how decimal fractions could be used but also advocated that a decimal system should be used for weights and measures and for coinage.

VI.11 René Descartes



b. La Haye (now “Descartes”), France, 1596; d. Stockholm, 1650
Algebra; geometry; analytic geometry; foundations of mathematics

In 1637 Descartes published *La Géométrie* as an “essay” appended to his philosophical treatise *Discours de la Méthode*. It remained his only mathematical publication. No single early modern text shaped the development of mathematics between 1650 and 1700 as strongly as *La Géométrie*. It was the founding text of analytic geometry and it paved the way for the merging of algebra and geometry that made possible the development of the integral and differential calculus about fifty years later.

Descartes was educated at the Jesuit College at La Flèche. He spent his life mostly outside France, traveling through Europe in his early twenties and living in the Netherlands from 1628 until 1649; he then left for Sweden, invited by Queen Christina to her court. From an early age his interest in mathematics was tightly linked to his primary philosophical preoccupation: the

certainty of knowledge. In a letter of 1619 he sketched a method, clearly inspired by arithmetic and geometry, for solving all problems in natural philosophy. Shortly afterward, his ideas grew into a passionate conviction that he could and should develop a philosophy along these problem-solving and mathematics-inspired lines. *La Géométrie* grew out of the mathematical part of his philosophical program; it was not a textbook on analytic geometry. Descartes offered little in the way of general principles, explaining his ideas by means of examples.

Descartes used a classical problem, Pappus’s problem, for explaining coordinates and equations of curves, and showed that the defining property of a curve could be written as an equation. He introduced coordinates x and y , using oblique as well as rectangular coordinate axes, which he always adjusted to the problem at hand. He also introduced the now very common usage of employing x , y , and z for unknowns and a , b , and c for indeterminate fixed quantities.

For Descartes, a geometrical problem required a geometrical answer. The equation was at best an algebraic reformulation of the problem; the answer had to be a construction of the curve or of individual points. If, as in the particular case of Pappus’s problem in four lines, the equation was quadratic, then for any fixed value of y the x -coordinate was a root of a quadratic equation. Earlier in the book Descartes had shown how such a root could be constructed (using ruler and compass). Thus, the curve could be constructed “pointwise” by choosing a series of values for y and constructing the corresponding x s and points on the curve. Pointwise construction could not provide the whole curve. Therefore in Pappus’s problem Descartes used the equation to show that the solution curves were conic sections, and explained how to determine the nature of the conic, the location of its axes, and the values of its parameters. This was an impressive result; it was, in fact, the first classification of an algebraically defined class of curves.

A further influential result in *La Géométrie*, and the one of which Descartes himself said he was most proud, was his method to determine the normal (and thus also the tangent) at a given point on a curve with a given equation. It was a pre-calculus forerunner of differentiation.

There are three important differences between how Descartes treated curves and their equations and how they are treated in modern analytic geometry: he employed oblique as well as rectangular axes; he did

not consider the equation as defining a curve—rather it represented a problem, namely to construct the curve itself, as well as its axes, tangents, etc.; and he did not consider the plane itself as a collection of points characterized by pairs of real numbers—for him the x s and y s were not dimensionless numbers but the lengths of line segments. (The term “Cartesian plane” for \mathbb{R}^2 is therefore anachronistic.)

Descartes supposed (too optimistically) that his procedures could be extended to polynomial equations of any degree (usually connected to Pappus’s problem with more than four lines) and that therefore he had shown how, in principle, all geometrical construction problems could be solved. For higher-order constructions he needed new algebraic techniques. The relevant section in *La Géométrie* constituted the first general theory of polynomial equations and their roots. It contained his “sign rule” about the number of positive and negative roots of a polynomial, various transformation rules, and methods to check equations for reducibility. He gave no proofs; his results were based on a conviction that polynomials could essentially be written as products of linear factors $x - x_i$, in which the roots x_i could be positive, negative, or “imaginary.”

It appears, then, that analytic geometry was not the primary goal of *La Géométrie*. Rather, its aim was to provide a universal method for solving geometrical problems, and to do so Descartes had to answer two urgent methodological questions. The first was how to solve geometrical problems not constructible by ruler and compass, and the second was how to use algebra as an analytic, i.e., solution-finding, tool in geometry.

For the first of these, Descartes allowed successively more complicated curves as means of construction. It was Descartes’s conviction that algebra, through the equations of these curves, could guide him to choose, among all such construction curves, the most appropriate for the problem, in particular the simplest, i.e., that of lowest degree.

The second question addressed serious conceptual difficulties that were felt at the time about using algebra in geometry. The transfer of algebraic operations to geometry was indeed problematic because multiplication in geometry was generally interpreted dimensionally: for example, a product of two lengths had to represent an area, and a product of three a volume. But until then algebra had dealt mostly with numbers and had routinely used products of more than three factors. Thus, a consistent and unrestricted geometrical interpretation of the operations of algebra was

needed. Descartes did indeed provide such a reinterpretation. He introduced a unit line segment in such a way that multiplication no longer raised the dimension and inhomogeneous terms could be allowed in equations.

By 1637 he had given up on his earlier attempts to link philosophy and mathematics. Yet the preoccupation with certainty remained. As his concept of construction involved the use of curves, he had to consider which curves could be understood by the human mind with sufficient clarity to be acceptable in geometry. His answer was that all algebraic curves were acceptable (he called these “geometrical curves”) and all others were not (these he called “mechanical”). Few seventeenth-century mathematicians followed Descartes in this strict demarcation of geometry. This is typical of the reception of Descartes’s *La Géométrie*: the philosophical and methodological aspects of the book were largely ignored by his mathematical readers, but the technical mathematical aspects were eagerly accepted and used.

Further Reading

- Bos, H.J.M. 2001. *Redefining Geometrical Exactness: Descartes’ Transformation of the Early Modern Concept of Construction*. New York: Springer.
- Cottingham, J., ed. 1992. *The Cambridge Companion to Descartes*. Cambridge: Cambridge University Press.
- Shea, W. R. 1991. *The Magic of Numbers and Motion: The Scientific Career of René Descartes*. Canton, MA: Watson Publishing.

Henk J. M. Bos

VI.12 Pierre Fermat

b. Beaumont-de-Lomagne, France, 1607; d. Castres, France, 1665
Number theory; probability theory; variational principles; quadrature; geometry

Fermat, who spent his life as a magistrate in the south of France, contributed decisively to most of the mathematical subjects of his time: from quadrature to optics, from geometry to number theory. Very little is known about his early life—even the date of his birth is uncertain—but by 1629 he had close contacts with VIÈTE’s [VI.9] scientific heirs in Bordeaux. His work displays a thorough knowledge of ancient as well as contemporary mathematics and he exchanged problems and mathematical information by correspondence with, among others, RENÉ DESCARTES [VI.11], Gilles Personne de Roberval, Marin Mersenne, Bernard Frenicle, John Wallis, and Christiaan Huygens.

A crucial early-modern topic was the use of algebra to solve geometric problems. Viète and other algebraists before him had used equations in a single unknown to rewrite and solve “determinate” problems (problems admitting a finite number of solutions). In his manuscript *Ad Locos Planos et Solidos Isagoge*, which circulated in Paris in 1637 (the same year as Descartes’s *La Géométrie*), Fermat presented a general way of handling and solving indeterminate problems associated with constructions of loci: that is, of sets of points (usually curves) defined by some constraints. He identified the points of such loci by two coordinates linked by an equation (although he chose a different way of taking coordinates from the usual modern x and y coordinates). Moreover, Fermat gave the standard forms of the corresponding equation when the locus to be found was a line, a parabola, an ellipse, etc.

Fermat also used algebraic analysis to solve problems of extrema, including finding the tangent or the normal to a curve at a given point, and determining centers of gravity. His method relies on the principle that a certain algebraic expression takes on the same values twice near the extremum. Although the procedure is purely algebraic, his successors tended to interpret it from a differential perspective, thereby making his work an apparent precursor of the calculus. Fermat applied the method to a variety of problems, including (within the framework of a controversy with Descartes’s followers around 1660) a proof of the law of refraction in optics. Basing his analysis on the principle that “nature acts in the shortest time,” Fermat was able to express the problem as one of extrema and to solve it with his method. The problem of refraction was one of the first complex physical problems to be treated in a thoroughly mathematical way, and Fermat’s approach later led to VARIATIONAL METHODS [III.96].

However, Fermat also showed a perfect mastery of more classical, for instance Archimedean, techniques, which he used when dealing with other types of geometrical questions such as quadrature.

Such versatility also appears in Fermat’s work on numbers. On the one hand, he was happy to apply his algebraic approach to Diophantine analysis in order to obtain solutions for cases previously thought to be insoluble, or to derive new solutions from ones already known. On the other hand, he advocated a theoretical study of the integers, for which the currently available algebraic theory of equations was not sufficient. For example, he gave general properties of the divisors of

numbers of the form $a^n \pm 1$ (among them his now celebrated LITTLE THEOREM [III.60]) and of $x^2 + Ny^2$ for various N . He invented the method of infinite descent specifically to deal with problems concerning integers. He used this method, which relies on the impossibility of constructing an infinite strictly decreasing sequence of integers, to prove that $a^4 - b^4 = c^2$ has no nontrivial integer solutions. This is a particular case of his famous LAST THEOREM [V.12], which Fermat only stated in the margins of one of his books: $a^n + b^n = c^n$ has no nontrivial integer solutions for $n > 2$. The first proof of the general case was given by Andrew Wiles in 1995.

In 1654 Fermat exchanged letters with PASCAL [VI.13] on the idea of a “fair game” and on the redistribution of the stakes if a game is interrupted before its end. These letters introduced important concepts in probability, including expected value and conditional probability.

Further Reading

- Cifoletti, G. 1990. *La Méthode de Fermat, Son Statut et Sa Diffusion*. Société d’Histoire des Sciences et des Techniques. Paris: Belin.
- Goldstein, C. 1995. *Un Théorème de Fermat et Ses Lecteurs*. Saint-Denis: Presses Universitaires de Vincennes.
- Mahoney, M. 1994. *The Mathematical Career of Pierre de Fermat (1601–1665)*, second revised edn. Princeton, NJ: Princeton University Press.

Catherine Goldstein

VI.13 Blaise Pascal

b. Clermont-Ferrand, France, 1623; d. Paris, 1662

Scientist and theologian

Pascal was the first to make a systematic study of the arithmetical triangle which now bears his name; although the triangle itself is found earlier, notably in the work of the Chinese mathematician Zhu Shijie (1303). “Pascal’s triangle”

$$\begin{array}{ccccccc}
 & & & & 1 & & & \\
 & & & & 1 & & 1 & \\
 & & & 1 & & 2 & & 1 \\
 & & 1 & & 3 & & 3 & & 1 \\
 & 1 & & 4 & & 6 & & 4 & & 1 \\
 & & \cdot & & \cdot & & \cdot & & \cdot & & \cdot
 \end{array}$$

a triangular array in which each number is the sum of the two immediately above it, provides a geometrical arrangement of the binomial coefficients $\binom{n}{k}$, with $\binom{n}{k}$

PUP: this has been double-checked and ‘c²’ is indeed correct.

appearing as the $(k + 1)$ st element in the $(n + 1)$ st row. Here $\binom{n}{k}$ is, as usual, the number of subsets of size k in a set of size n , so that

$$\binom{n}{k} = \frac{n!}{k!(n - k)!}.$$

The number $\binom{n}{k}$ is also the coefficient of $a^k b^{n-k}$ in the binomial expansion of $(a + b)^n$ for any integer $n \geq 0$ and $0 \leq k \leq n$. In his *Traité du Triangle Arithmétique* (printed in 1654 but not distributed until 1665) Pascal was the first to connect binomial coefficients with the combinatorial coefficients that arise in probability. The *Traité* is famous too for its explicit statement of the principle of mathematical induction.

Pascal is also known for a theorem in projective geometry (given an arbitrary hexagon inscribed in any conic section, if the three pairs of opposite sides are continued until they meet, then the three points of intersection lie on a straight line) (1640); and for a two-function (addition and subtraction) mechanical calculating machine (1645).

VI.14 Isaac Newton



b. Woolsthorpe, England, 1642; d. London, 1727
Calculus; algebra; geometry; mechanics; optics;
mathematical astronomy

Newton entered Trinity College, Cambridge, in 1661, and it was in Cambridge that he spent most of his formative years, first as a student, then as a Fellow, and then, from 1669, as Lucasian Professor of Mathematics. His election to the Lucasian Chair was engineered by his mentor Isaac Barrow, a talented mathematician and theologian who was the first to hold the prestigious chair. In 1696 Newton moved to London to take

up the post of Warden of the Mint. He resigned his professorship in 1702.

It appears that Newton's interest in mathematics began in 1664 when, embarking on a course of self-instruction, he read VIÈTE's [VI.9] works (1646), Oughtred's *Clavis Mathematicae* (1631), DESCARTES's [VI.11] *La Géométrie* (1637), and Wallis's *Arithmetica Infinitorum* (1656). From Descartes, Newton learned how useful it could be to relate algebra to geometry, since plane curves could be represented by algebraic equations in two unknowns. Descartes had, however, imposed strict limitations on the class of curves allowed in *La Géométrie*: "geometrical" (i.e., algebraic) curves were admitted but "mechanical" (i.e., transcendental) ones were not. In common with many of his contemporaries, Newton felt that such limitations ought to be overcome and that a "new analysis" capable of dealing with mechanical curves ought to be possible. He found the answer in infinite series.

Newton had learned how to deal with infinite series from Wallis's work, and it was while elaborating one of Wallis's techniques that, in the winter of 1664, he obtained his first great mathematical discovery: the binomial theorem for fractional powers. This provided him with a method for expanding into power series a large class of "curves," including transcendental curves, which could now be given an "analytical" representation to which the rules of algebra could be applied. Termwise application of the relation (which he knew from Wallis and which is expressed in familiar Leibnizian notation as $\int x^n dx = x^{n+1}/(n + 1)$) allowed him to "square" a variety of curves when they were expanded as a power series. (In the seventeenth century, squaring a curvilinear figure meant finding a square the area of which is equal to that of the curvilinear figure.)

A few months later, Newton, with extraordinary insight, realized that most of the problems dealt with by his contemporaries could be reduced to two classes: problems in which one is required to find the tangent to a curve, and problems in which one is required to find the area subtended by a curve. He conceived geometrical magnitudes as being generated by continuous motion. For example, the motion of a point generates a line and the motion of a line generates a surface. These he called "fluents," while the instantaneous rate of flow he called the "fluxion." Basing his intuitions on kinematical models, he formulated a version of what is known today as THE FUNDAMENTAL THEOREM OF CALCULUS [I.3 §5.5]. Namely, he proved that tangent and

area problems are inverses of each other. In modern terms, Newton was able to reduce quadrature problems (i.e., calculating curvilinear areas) to the search for primitive functions (indefinite integrals). He built “catalogues of curves” (tables of integrals), deploying techniques equivalent to substitution of variables and integration by parts. He developed an efficient algorithm that allowed him to tackle both the direct (differential) and inverse (integral) methods of fluxions. He was able to calculate the tangent to and curvature of any known curve, and perform integrations of many classes of (what we now call) ordinary differential equations. Such mathematical tools allowed him to explore the properties of cubics, and he classified seventy-two different species of them. His results on series and on the direct and inverse methods of fluxions were published in *De Quadratura Curvarum* and those on cubics in the *Enumeratio Linearum Tertii Ordinis*, both works appearing in 1704 as appendices to *Opticks*. His *Arithmetica Universalis*, the text in which he collected together his lectures on algebra, appeared in 1707.

Before 1704, Newton, displaying his characteristic reluctance to publish, had divulged his discoveries on the fluxional method through letters and manuscripts rather than in printed form. In the meantime LEIBNIZ [VI.15], later than Newton but independently, had also discovered the differential and integral calculus, and had printed it as early as 1684–86. Newton was convinced that Leibniz had stolen the idea from him, and from 1699 onward he engaged Leibniz in a bitter quarrel over priority.

In the early 1670s Newton began distancing himself from the modern symbolic style that had characterized his youthful researches. He turned to geometry in the hope of restoring a hidden geometric method of discovery: the “method of analysis,” known to the ancient Greeks. In fact, geometry dominates Newton’s masterpiece, the *Philosophiae Naturalis Principia Mathematica*. In this work, which appeared in 1687, Newton presented his theory of gravitation. Newton was convinced that the ancient method was superior to the modern symbolic one that he identified with Cartesian analysis. In his attempts to rediscover the method, he developed elements of projective geometry. (This sprang from the idea that the ancients were able to solve complex problems related to conics by using projective transformations.) An important result is his solution of Pappus’s locus problem, which appears in book I of the *Principia* (1687). Here he shows that a conic is the locus of points, the product of whose distances from two given lines

is proportional to the product of its distances from a third and fourth given line. He then applied projective transformations to determine the conic tangent to m given lines that passes through n given points, when $m + n = 5$.

The *Principia* contains a rich array of mathematical results. In book I Newton presents the “method of first and ultimate ratios,” in which he deploys geometric limit procedures in order to determine tangents, curvatures, and curvilinear areas, the latter containing the basic ingredients of what today is known as the RIEMANN INTEGRAL [I.3 §5.5]. He also shows that “ovals” are algebraically nonintegrable. In dealing with the so-called Kepler problem, Newton approximates the roots of $x - d \sin x = z$ (d and z given) by a technique equivalent to the NEWTON-RAPHSON METHOD [II.4 §2.3]. In book II he inaugurates VARIATIONAL METHODS [III.96] by tackling the problem of the solid with least resistance. And in book III, in dealing with cometary paths, he presents a method of interpolation which inspired research by mathematicians such as Stirling, Bessel, and GAUSS [VI.26]. In his masterpiece, Newton had shown how productive the application of mathematics to natural philosophy could be: most notably, his studies on the Moon’s motion, the precession of the equinoxes, and the tides were seminal in stimulating eighteenth-century perturbation theory.

Further Reading

- Newton, I. 1667–81. *The Mathematical Papers of Isaac Newton*, edited by D. T. Whiteside et al., eight volumes. Cambridge: Cambridge University Press.
- Pepper, J. 1988. Newton’s mathematical work. In *Let Newton Be! A New Perspective on His Life and Works*, edited by J. Fauvel, R. Flood, M. Shortland, and R. Wilson, pp. 63–79. Oxford: Oxford University Press.
- Whiteside, D. T. 1982. Newton the mathematician. In *Contemporary Newtonian Research*, edited by Z. Bechler, pp. 109–27. Dordrecht: Reidel. (Reprinted, 1996, in *Newton. A Critical Norton Edition*, edited by I. B. Cohen and R. S. Westfall, pp. 406–13. New York/London: W. W. Norton & Co.)

Niccolò Guicciardini



b. Leipzig, Germany, 1646; d. Hanover, Germany, 1716
Calculus; theory of linear equations and elimination theory; logic

Renowned among mathematicians for his invention of the calculus, Leibniz was a universal thinker who graduated in law and was self-taught in mathematics. In 1676 he became counselor and librarian in Hanover for the Duke Johann Friedrich of Braunschweig-Lüneburg, holding this position until the end of his life. Besides mathematics he occupied himself with technical, historiographical, political, religious, and philosophical questions. His philosophy distinguished between two areas of reality: the world of appearances and the world of substances. It was in developing his philosophy that he was led to declare that the real world is “the best of all possible worlds.” In 1700 he was appointed first president of the newly founded Brandenburg Society of Sciences established in Berlin.

Most of his mathematical ideas and writings were not published during his lifetime, and consequently many of his results were rediscovered many years later. About a fifth of his mathematical papers have now been published. He was always more interested in general or even universal methods than in technical details, using analogy and inductive reasoning to develop the art of invention. For the same reason he became a key creator of mathematical notation: he knew how much a suitable notation could facilitate mathematical discoveries.

One of Leibniz’s earliest mathematical works was a treatise on infinitesimal geometry (written in 1675–76 but not published until 1993). In it he used his “quanta” concept of the infinite. In Leibniz’s eyes the actual infinite as well as indivisibles, in the strictest sense of the word, were not quantities and therefore not mathematical entities: hence, he used the notions “infinitely small”

and “infinitely large.” These denoted, it is true, variable quantities, but nevertheless they were quantities of a sort, so they could be handled by mathematics. Among the results in this treatise is a rigorous proof, in the style of ARCHIMEDES [VI.3], of the existence of (what is today known as) the RIEMANN INTEGRAL [I.3 §5.5] of continuous functions, which is based on intermediary values of the function within subintervals. Only a few of these results were actually published by Leibniz, and even these mainly without proof: in 1682 the alternating series for $\pi/4$; in 1691 some further results. In 1713 he communicated his alternating series test in a private letter to JOHANN BERNOULLI [VI.18].

The year 1675 was also the year in which Leibniz invented his version of the differential and integral calculus, although its publication did not begin until 1684. His calculus was based on the key concept of a variable (quantity) ranging over a sequence of values infinitely close to each other, with the differential, the difference between two successive values in the sequence, being itself a variable that could be manipulated in the usual manner. Differentiation was represented by the operator “d”, which assigned variables to variables. For example, if x is a line of variable length, then dx is a very short line, also of variable length. Integration meant summation. His notation (d and \int) is still used today. He deduced the standard differentiation rules (the chain rule, the product rule, etc.) and successfully applied his calculus to the differentiation of families of curves, to differentiation under the integral sign, and to various types of differential equations.

Leibniz considered “combinatorial art” as a general qualitative science, which did not coincide with modern combinatorial analysis but included combinatorics and algebra: Leibniz considered it as “the inventive part of logic.” Here he found the Girard formula for the representation of sums of powers of roots of equations by means of elementary symmetric functions, and the so-called Waring formulas by which polynomial symmetric functions are reduced to power sums (these were rediscovered by WARING [VI.21] in 1762). He invented double and multiple indices in order to solve systems of linear equations and problems of elimination theory. Between 1678 and 1713 he laid the foundations for the theory of DETERMINANTS [III.15]. The method now known as Cramer’s rule, for solving simultaneous equations, which in modern terms is based on determinants, and which Cramer published in 1750, was in fact found in 1684 by Leibniz (but again not published by him). He also stated (without proof) several theorems

in the theory of linear equations and elimination theory now attributed to EULER [VI.19], LAPLACE [VI.23], and SYLVESTER [VI.42].

Among Leibniz's other mathematical interests was additive number theory. In 1673 he found a recursion formula for the number of tripartitions of a natural number (published in 1976) and discovered further rules of recurrence now attributed to Euler. He also developed a formalism for a positional calculus (*calculus situs*) in order to express positions in space: if the definitions of figures are completely expressed by this calculus, all of their properties can then be found by this calculus. This is closely linked with the modern notions of geometry and topology.

Leibniz was one of the pioneers of actuarial theory. Using mathematical models of human life he calculated the purchase price of life annuities both for single persons and for groups of men, and he applied such considerations to the liquidation of a state's indebtedness.

From the very beginning of his scientific career Leibniz was deeply interested in logic. He conceived of a general science: that is, of an art of inventing and of judging all sciences by means of sufficient data and a suitable universal language or writing. Yet, his "characteristica universalis" and the ensuing logical calculi remained fragmentary projects. His "calculus ratiocinator" was meant to be a formalized deduction of truth. Given that Leibniz was interested in formalizing calculations, it is not surprising that he also constructed the first four-function calculating machine. In constructing this machine he invented a new technical device, which he developed in two different versions: the so-called pinwheel (before 1676) and the stepped drum (from 1693 or earlier).

Further Reading

Leibniz, G. W. 1990-. *Sämtliche Schriften und Briefe, Reihe 7 Mathematische Schriften*, four volumes (so far). Berlin: Akademie.

Eberhard Knobloch

VI.16 Brook Taylor

b. Edmonton, Middlesex, England, 1685; d. London, 1731

Secretary of the Royal Society (1714–18)

Taylor was not the first to discover the theorem that bears his name (James Gregory found the theorem in

1671), but he was the first to publish it and the first to appreciate its significance and applicability. The theorem, which states that any function that satisfies certain conditions can be expressed as (what is now known as) a Taylor series, was published in Taylor's *Methodus Incrementorum Directa et Inversa* (1715). In the *Methodus* Taylor gave the series as

$$f(x+h) = f(x) + \frac{f'(x)}{1!}h + \frac{f''(x)}{2!}h^2 + \frac{f'''(x)}{3!}h^3 + \dots$$

(as it would appear in modern notation). Although Taylor did not attend to questions of rigor—there is no consideration of convergence, of the remainder term, or of the validity of expressing a function by such a series—his derivation of the series was not out of line with the standards of its day. Taylor used the theorem for approximating the roots of equations and for solving differential equations. Although he was aware of its use for expanding functions into series, he does not appear to have fully appreciated its significance in this respect.

Taylor is also noted for his contribution to the problem of the vibrating string (discussed in the *Methodus* and in earlier papers) and for a book on the theory of linear perspective (1715).

VI.17 Christian Goldbach

b. Königsberg (now Kaliningrad, Russia), 1690; d. Moscow, 1764

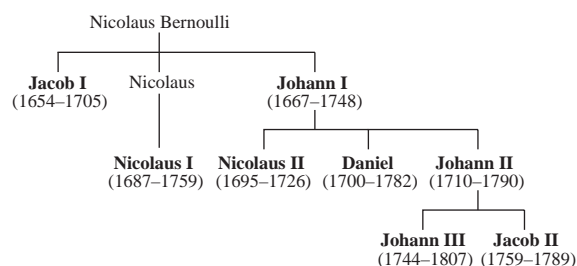
Professor of Mathematics, Imperial Academy of Sciences, Saint Petersburg (1725–28); tutor to Tsarevitch Peter II, Moscow (1728–30); corresponding secretary and administrator, Imperial Academy of Sciences, Saint Petersburg (1732–42); Ministry of Foreign Affairs (1742–64)

Goldbach is remembered today for the conjecture that bears his name: that every even number greater than 2 is a sum of two primes. This was first stated by EULER [VI.19] in 1742 in a letter to Goldbach in response to the earlier proposal by Goldbach that every number greater than 2 is a sum of three primes (Goldbach considered 1 as a prime number). Goldbach's conjecture, together with the weaker conjecture that every odd number is either prime or the sum of three primes, was first published by WARING [VI.21] in 1770 but without attribution. Both conjectures remain unsolved. However, Vinogradov proved that every sufficiently large odd number is the sum of three primes: see PROBLEMS AND RESULTS IN ADDITIVE NUMBER THEORY [V.30].

PUP: this has been double-checked and is correct.

PUP: this has been checked and there are still only four volumes at present.

VI.18 The Bernoullis



All born in Basel, Switzerland, apart from Daniel (Groningen, the Netherlands). All died in Basel, apart from Jacob II, Nicolaus II (both Saint Petersburg, Russia), and Johann III (Berlin). (The two members of the family not in bold text were not mathematicians.)

The Bernoullis played a remarkable role in the development of mathematics during the Enlightenment. Indeed such was the family's importance that in 1715 LEIBNIZ [VI.15] coined the term “bernoullizare” to describe the activity of doing mathematics. Altogether eight members of the family devoted themselves to the mathematical sciences (including physics, especially mechanics and fluid mechanics), and from 1687 to 1790 the mathematics chair of the university of Basel was occupied by successive members of the family: first Jacob (1687–1705), then his brother Johann (1705–48), and finally Johann's son, Johann II (1748–90). Throughout the eighteenth century Bernoullis were members of the Paris Academy of Sciences and individually they won prestigious prizes on many occasions. The same was true of the academies in Berlin, Saint Petersburg, and several others.

The family goes back to a line of Calvinist merchants who fled the Spanish Netherlands. The first Bernoulli to settle in Basel was Jacob, a druggist, who became a citizen in 1622. His grandson, Jacob I, studied philosophy and theology in Basel before turning to mathematics, against the will of his father. This was to be a typical pattern in the family: many of the Bernoullis studied mathematics despite pressure from the family to make a career in other areas (such as medicine or law). Having received a licentiate in theology in 1676, Jacob undertook an educational journey which took him first to France, then to the Netherlands, and finally to England. And it was through his encounters with Nicolas Malebranche, Jan Hudde, and others that he became acquainted with Cartesianism and its most eminent representatives. In 1677 he started his

diary, *Meditationes*, in which he wrote down many of his mathematical insights and thoughts.

Having obtained the chair of mathematics in Basel, Jacob studied Leibniz's early memoirs on differential calculus, whose power he was the first, together with his younger brother Johann I, to recognize. In a paper on the curve of constant descent, published in the Leipzig *Acta Eruditorum* in 1690, Jacob was the first to use the term “integral” in its present mathematical sense. From then on he showed his mastery of Leibnizian methods in his study of curves, including, among others, the catenary, the form of a bent elastic beam, the form of a sail inflated by the wind, and the parabolic and logarithmic spirals. He also solved the differential equation $y' = p(x)y + q(x)y^n$, which is now named after him. However he is best remembered for his *Ars Conjectandi* (1713), which was published posthumously with a short foreword by his nephew, Nicolaus I. It contains an attempt to give a sound mathematical treatment of a commonsense principle already appealed to by CARDANO [VI.7] and Halley: if an experiment is repeated a large number of times, then the relative frequency with which an event occurs will roughly equal the probability of the event. Bernoulli's theorem, known since POISSON [VI.27] as the (WEAK) LAW OF LARGE NUMBERS [III.73 §4], establishes a first link between the theories of probability and statistics. In the same book, Bernoulli also introduced the sequence B_0, B_1, \dots of rational numbers that now bears his name, which can be defined as the coefficients of $t^k/k!$ in the power-series expansion

$$\frac{t}{e^t - 1} = \sum_{k=0}^{\infty} B_k \frac{t^k}{k!}.$$

Jacob computed these numbers up to B_{10} .

Johann, who had to study medicine before he could devote himself to mathematics, got his first mathematical training from his brother Jacob, with whom he developed numerous applications of the new Leibnizian calculus to mechanics. An academic peregrination brought him to Paris in 1691–92, where he gave private lessons to Guillaume de l'Hôpital. These lessons are the basis of l'Hôpital's famous *Analyse des Infinites Petits* (1696). This textbook, the first on calculus, contains l'Hôpital's rule, which Johann had communicated by letter to his student. In 1695, Johann left Basel for Groningen to take up a professorship in mathematics.

With the growing visibility of Johann's work, the friendly collaboration between the two brothers Jacob

and Johann transformed into an endless round of controversies, priority disputes, and public accusations. They engaged in heated struggles concerning the solution of the brachistochrone (curve of fastest descent) problem, and a complicated isoperimetric problem that involved minimizing the area enclosed by a curve of fixed length. Eventually, these bitter quarrels led to an interesting mathematical outcome: the creation of the CALCULUS OF VARIATIONS [III.96]. After Jacob's death, Johann took over the mathematics chair in Basel, where he taught until the end of his life, attracting students from all over Europe, including EULER [VI.19].

Johann's most important achievement in mathematics is the development of the integral calculus. He developed a general theory for the integration of rational functions and new methods for the solution of differential equations. He also extended the infinitesimal calculus to handle EXPONENTIAL FUNCTIONS [III.25].

Johann's correspondence with Leibniz (which spans approximately twenty-five years) can be viewed as a laboratory of mathematical invention and debate. The priority dispute that ensued when NEWTON [VI.14] accused Leibniz of having stolen the calculus from him also involved Johann, who fought on Leibniz's side. With each camp defying the other with difficult problems, Johann had the opportunity to create, with his son Nicolaus II, the theory of orthogonal trajectories of families of curves. Johann was also a towering figure in the origins of analytical mechanics and in mathematical physics, where, among other things, he made notable contributions to the study of central forces, to navigational theories, and to the question of the principles of statics.

Nicolaus I studied mathematics at the university of Basel with his uncle Jacob before taking a degree of Doctor of Jurisprudence (1709). He was the professor of mathematics in Padua, occupying the chair once held by Galileo, and he later held the professorship of logic in Basel. His main interests in mathematics were infinite series and the applications of probability theory to questions of law. He formulated, in 1713, the notorious Saint Petersburg paradox, which originated with a gambling game. Suppose that Peter is tossing a fair coin, and he will give Paul one ducat if the coin turns up heads on the first toss, two ducats if it shows heads for the first time on the second toss, and in general 2^{n-1} ducats if the coin turns up heads for the first time on the n th toss. The standard calculation shows that the value of Paul's expectation ($E = \frac{1}{2}1 + \frac{1}{4}2 + \frac{1}{8}4 + \cdots + \frac{1}{2^n}2^{n-1} + \cdots$) is infinitely

great. Nevertheless no "fairly reasonable man" would be willing to pay even a moderately high price to purchase Paul's prospects. The result of the mathematical analysis clearly affronted common sense; this was the paradox. Nicolaus's cousin, Daniel, discussed the problem while he was staying in Saint Petersburg (hence the name given to the paradox). His strategy was to distinguish two senses of expectation, one mathematical and the other moral. The latter was to take into account the individual characteristics of the risk taker (his wealth, for instance).

Although primarily a physicist and author of the famous *Hydrodynamica* (1738), Daniel obtained a solution of the Riccati equation, $y' = r(x) + p(x)y + q(x)y^2$, and engaged in the problem of the vibrating string.

Otto Spiess, from Basel, started to publish the complete edition of the works and correspondence of the Bernoullis in 1955. The project continues.

Further Reading

Cramer, G., ed. 1967. *Jacobi Bernoulli, Basileensis, Opera*, two volumes. Brussels: Editions Culture et Civilization. (Originally published in Geneva in 1744.)

—. 1968. *Opera Omnia Johannis Bernoulli*, four volumes. Hildesheim: Georg Olms. (Originally published in Lausanne and Geneva in 1742.)

Spiess, O., ed. 1955-. *The Collected Scientific Papers of the Mathematicians and Physicists of the Bernoulli Family*. Basel: Birkhäuser.

PUP: this is indeed an ongoing series of publications.

Jeanne Peiffer

VI.19 Leonhard Euler



b. Basel, Switzerland, 1707; d. Saint Petersburg, Russia, 1783
Analysis; series; rational mechanics; number theory;
music theory; mathematical astronomy;
calculus of variations; differential equations

Euler was one of the most influential and prolific mathematicians in history. His first publication was a 1726 paper on mechanics, and his last was a collection published in 1862, seventy-nine years after his death. There are over eight hundred papers bearing his name, about three hundred of them appearing posthumously, and more than twenty books. His *Opera Omnia* fill over eighty volumes.

In number theory, Euler introduced the Euler phi function, $\phi(n)$, to denote the number of positive integers less than n and relatively prime to n , and proved the FERMAT-EULER THEOREM [III.60] that n divides $a^{\phi(n)} - 1$. He showed that the remainders relatively prime to n form what we now call a group under multiplication and he expanded the theory of quadratic and higher-order residues. He proved FERMAT'S LAST THEOREM [V.12] for $n = 3$. He stated that any real polynomial of degree n is a product of real and quadratic factors and has n complex roots, but was unable to give complete proofs. He was the first to use GENERATING FUNCTIONS [IV.18 §§2.4, 3] when he gave a generating function for Naudé's partition problem: the question of how many different ways a given integer can be written as a sum of positive integers. He introduced the function $\sigma(n)$, the sum of the divisors of an integer n , and used this function to increase the number of known pairs of amicable numbers (a pair m, n of numbers is called *amicable* if the sum of the proper divisors of m equals n , and vice versa) from 3 to over 100. He showed that any prime number of the form $4n + 1$ is the sum of two rational squares. LAGRANGE [VI.22] later improved this result to show that such numbers are the sum of two integer squares. Euler factored the fifth Fermat number, $F_5 = 2^{2^5} + 1$, thus refuting FERMAT'S [VI.12] conjecture that all integers of the form $F_n = 2^{2^n} + 1$ were prime. He made extensive studies of the binary quadratic forms $x^2 + y^2$, $x^2 + ny^2$, and $mx^2 + ny^2$, and proved a form of the LAW OF QUADRATIC RECIPROCITY [V.31].

Euler was the first to use analytic methods in number theory. In the 1730s he calculated to several decimal places the so-called Euler-Mascheroni constant

$$\gamma = \lim_{n \rightarrow \infty} \left[\left(\sum_{k=1}^n \frac{1}{k} \right) - \log n \right]$$

and discovered many of its properties. Mascheroni added to those properties in the 1790s. Euler also discovered the sum-product formula for what we now call the Riemann zeta function,

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{p \text{ prime}} \frac{1}{1 - p^{-s}},$$

and he evaluated the function for positive even values of s .

In analysis, besides defining the modern calculus curriculum, Euler was the first to take a systematic approach to the solution of differential equations and to problems of the CALCULUS OF VARIATIONS [III.96]. He discovered a differential equation sometimes called the "Euler necessary condition" and sometimes called the "Euler-Lagrange equation." The equation tells us that if J is defined by the integral equation $J = \int_a^b f(x, y, y') dx$, then a function $y(x)$ that maximizes or minimizes J will satisfy the differential equation

$$\frac{\partial f}{\partial y} - \frac{d}{dx} \left(\frac{\partial f}{\partial y'} \right) = 0.$$

Euler apparently thought that the condition was also sufficient. Very early in his career, he pioneered the use of the integrating factor in differential equations, though the almost simultaneous published solution of Clairaut was more complete and more widely read, so credit for this innovation usually falls to Clairaut. He also did the first work using what are now called FOURIER SERIES [III.27] and LAPLACE TRANSFORMS [III.93], more than a generation before LAPLACE [VI.23] or FOURIER [VI.25] began doing mathematics, though they took the fields much farther than Euler had.

Much of Euler's best work involved series. His first widely acclaimed result was when he solved one of the best-known problems of his age, the seventy-year-old "Basel problem." The problem was to evaluate the sum of the reciprocals of the square integers, or $\zeta(2)$. Euler showed that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

(For a sketch of a proof, see π [III.72].)

He developed the Euler-Maclaurin series to strengthen the relationships between series and integrals. The existence of the Euler-Mascheroni constant followed from these researches. Using techniques he called "interpolation of series," he developed the GAMMA FUNCTION [III.31] and the beta function. He developed the first extensive theory of CONTINUED FRACTIONS

[III.22], and derived series for the accurate and efficient calculation of LOGARITHMS [III.25 §4] and trigonometric tables, often to more than twenty decimal places.

He was the first to do calculus with complex numbers and to investigate logarithms of negative and complex numbers. This research led to a long and bitter controversy with D'ALEMBERT [VI.20].

Euler was not the first to prove that $e^{i\theta} = \cos \theta + i \sin \theta$ or to know that $e^{\pi i} = -1$, but he made so much more use of these facts than any of his predecessors that this last formula is generally known as Euler's identity.

He is regarded as a pioneer in topology and graph theory for his necessary condition for a graph to have an Euler path, the so-called Königsburg bridge problem. This is to determine whether or not a graph has a path that traverses every edge exactly once. He also discovered and gave a flawed proof that, for a polyhedron "bounded by planes," Euler's words for what we now call "convex," $V - E + F = 2$, where V is the number of vertices, E is the number of edges, and F is the number of faces. (For details about the flaws in Euler's proof, see Richeson and Francese (2006).)

Euler proved a form of the general addition theorem for elliptic integrals and gave a complete classification of elastic curves. At the command of his king, Frederick the Great of Prussia, he studied hydraulics, designed pumps and fountains, and evaluated the probabilities and combinatorics involved in the state lotteries.

In a triangle, the line on which the orthocenter, the centroid, and the circumcenter lie is the Euler line. The Euler method is an algorithm for giving numerical solutions to differential equations. The EULER DIFFERENTIAL EQUATION [III.23] is the partial differential equation that describes continuity of fluid flow.

Euler tried to use lunar and planetary theory to solve the problem of finding longitude at sea. In studying the orbit of a comet, he made the first steps in the statistics of observed data.

He left Switzerland in 1727 to work in the new academy of Peter the Great in Saint Petersburg. In 1741 he moved to Berlin and the academy of Frederick the Great, but returned to Saint Petersburg in 1766, after the ascension of Catherine the Great. He was blind for the last fifteen years of his life, during which time he nevertheless wrote over three hundred papers. He won the annual prize competition of the Paris Academy twelve times.

His series of calculus books, published in four volumes between 1755 and 1770, were the first success-

ful calculus textbooks. It was the climax of a complete series of mathematical textbooks, including arithmetic (1738), algebra (1770), and the *Introductio in Analysin Infinitorum* (1748), a textbook on the mathematics Euler thought was necessary to understand calculus.

In the two volumes of the *Mechanica* (1736), Euler gave the first calculus-based treatment of the mechanics of point masses. He followed this with another two-volume work, *Theoria Motus Corporum* (1765), on the motions of solid bodies, including rotations.

Other books include *Methodus Inveniendi* (1744), the first unified treatment of the calculus of variations, *Tentamen Novae Theoriae Musicae* (1739), on the physics of music and including the first use of logarithms in the theory of pitch, three different books on celestial mechanics and lunar theory, two on the theory of shipbuilding, three on optics, and one on ballistics.

Our modern notion that FUNCTIONS [I.2 §2.2] are a fundamental object in mathematics is due to Euler. Euler standardized the use of the symbols e , π , and i , as well as \sum for summations and Δ for finite differences.

His *Letters to a German Princess* in three volumes (1768–71) is regarded variously as the first work of popular science writing by a first-rate scientist and as an important work in the philosophy of science.

Laplace is reported to have advised, "Read Euler. Read Euler. He is the master of us all." The words are probably not those of Laplace, but the misattribution does not affect the quality of the advice.

Further Reading

- Bradley, R. E, and C. E. Sandifer, eds. 2007. *Leonhard Euler: Life, Work and Legacy*. Amsterdam: Elsevier.
- Dunham, W. 1999. *Euler: the Master of Us All*. Washington, DC: Mathematical Association of America.
- Euler, L. 1984. *Elements of Algebra*. New York: Springer. (Reprint of 1840 edition. London: Longman, Orme, and Co.)
- . 1988, 1990. *Introduction to Analysis of the Infinite*, books I and II, translated by J. Blanton. New York: Springer.
- . 2000. *Foundations of Differential Calculus*, translated by J. Blanton. New York: Springer.
- Richeson, D., and C. Francese. 2007. The flaw in Euler's proof of his polyhedral formula. *American Mathematical Monthly* 114(1):286–96.

Edward Sandifer

PUP: you said you wanted square brackets for the citations, and of course I'll do that if it's what you want, but with all the CRs having square brackets I have a definite preference for parentheses for the citations. Let me know what to do.

VI.20 Jean Le Rond d'Alembert

b. Paris, 1717; d. Paris, 1783

Algebra; infinitesimal calculus; rational mechanics; fluid mechanics; celestial mechanics; epistemology

D'Alembert spent his whole life in Paris, where he became one of the most influential members of the Académie Royale des Sciences and of the Académie Française. He became well-known as the scientific editor of the celebrated French *Encyclopédie*, the twenty-eight-volume work on which he collaborated with Denis Diderot, and for which he wrote most of the mathematical and many of the scientific articles.

As a student at the Jansenist Collège des Quatre-Nations, he followed the usual curriculum of grammar, rhetoric, and philosophy, the latter including some Cartesian science, a little mathematics, and much theology framed by the then burning debate about predestination, freedom, and grace. Disgusted with the permanent climate of controversy and the endless metaphysical discussions among his Jansenist teachers, d'Alembert decided, after attaining his diploma in law, to devote himself to his personal passion, "géométrie" (that is, mathematics).

D'Alembert's first communications to the French Académie were concerned with the analytic geometry of curves, the integral calculus, and fluid resistance, notably the problem of the deceleration and deflection of a disk entering a fluid, which was linked to the Cartesian explanation of refraction of light. He made a close reading of NEWTON's [VI.14] *Principia*, his commentary on passages of the first book showing his clear preference for analytical methods over the synthetic geometry of Newton.

D'Alembert's *Traité de Dynamique* (1743) made him famous in learned circles. He built up a systematic and rigorous theory of mechanics founded upon a short list of well-chosen principles— inertia, composition of motions (i.e., the addition of the effects of two forces or powers), and equilibrium—while at the same time trying to avoid metaphysical arguments. Most notably he proposed an important general principle, known today as "d'Alembert's principle," to simplify the investigation of constrained systems, such as the compound pendulum, vibrating rods, strings, rotating bodies, and even fluids, which he considered to be aggregates of parallel slices. The essential idea behind the principle was to reduce a problem in dynamics to one in statics, roughly speaking by introducing a fictitious force, the

"kinetic reaction," which was minus the mass times the acceleration. This allowed techniques from statics to be brought to bear on problems in dynamics.

His other books and memoirs were developments, some very innovative, in fluid theory, partial differential equations, celestial mechanics, algebra, and integral calculus. He devoted much thought to the use and status of imaginary numbers.

In his *Réflexions sur la Cause Générale des Vents* (1747) and his *Recherches sur le Calcul Intégral* (1748) he observed that numbers of the form $a + bi$ (where $i = \sqrt{-1}$) retain the same form when subjected to the usual operations (addition, subtraction, multiplication, division, and exponentiation). He proved that, for a real polynomial, imaginary roots always occur in conjugate pairs, and that even if a real polynomial has no real root, there is still always a complex root. However, his work was not rigorous—for example, he presupposed the existence of roots—and consequently he did not provide a proof of THE FUNDAMENTAL THEOREM OF ALGEBRA [V.15].

At the end of the 1740s there was a crisis in Newtonian science, with d'Alembert, Clairaut, and EULER [VI.19] each independently coming to the conclusion that Newton's theory of gravitation could not account for the motion of the Moon. In 1747 d'Alembert discussed various possibilities for solving the problem—an additional force, or a very irregular shape for the Moon, or some vortices between Earth and Moon—and produced a long study on celestial mechanics and planetary perturbations that has only recently been rediscovered and published (see d'Alembert 2002). By 1749 an improved mathematical analysis of the problem had shown that Newton's theory was correct. The rest of d'Alembert's extensive work on celestial mechanics was published in his *Recherches sur la Précession des Équinoxes* (1749), *Recherches sur Différents Points du Système du Monde* (1754–56), and in some of the eight volumes of his *Opuscules* (1761–83).

In 1747 d'Alembert presented a paper on the famous problem of vibrating strings, *Recherches sur la Courbe que Forme une Corde Tendue Mise en Vibration* (1749). This paper contained a solution of THE WAVE EQUATION [I.3 §5.4]. This was the first solution of a partial differential equation—partial differential equations were a new tool that he had already used in his 1747 *Réflexions sur la Cause Générale des Vents*. It led to a lengthy debate with Euler and DANIEL BERNOULLI [VI.18] about the possible form of the solutions and the general notion of function.

D'Alembert's work for the *Encyclopédie* (1751–65) and his efforts to find rigorous foundations for the sciences led him into the field of philosophy, where his main contributions concerned the classification of various sciences. He also worked on the study of cognition following the lines proposed by DESCARTES [VI.11], Locke, and Condillac.

Further Reading

- D'Alembert, J. le R. 2002. *Premiers Textes de Mécanique Céleste*, edited by M. Chapront. Paris: CNRS.
 Hankins, T. 1970. *Jean d'Alembert, Science and the Enlightenment*. Oxford: Oxford University Press.
 Michel, A., and M. Paty. 2002. *Analyse et Dynamique. Études sur l'Oeuvre de d'Alembert*. Laval, Québec: Les Presses de l'Université Laval.

Francois de Gandt



b. Turin, Italy, 1736; d. Paris, 1813
Number theory; algebra; analysis; classical and celestial mechanics

VI.21 Edward Waring

b. Shrewsbury, England, ca. 1735; d. Shrewsbury, 1798

Lucasian Professor of Mathematics, Cambridge (1760–98)

Waring, the leading British mathematician of the latter half of the eighteenth century, wrote several advanced but somewhat impenetrable analytical texts. His first work, *Miscellanea Analytica* (1762), is devoted to the theory of numbers and algebraic equations and contains many results which he revised and expanded in his *Meditationes Algebraicae* (1770). Included in the latter is the problem known today as Waring's problem (that every positive integer is the sum of not more than nine cubes, or the sum of not more than nineteen fourth powers, and so on, with a fixed number of summands depending on the exponent), which was solved by HILBERT [VI.63] in the affirmative in 1909 and which gave rise to important work by HARDY [VI.73] and LITTLEWOOD [VI.79] in the 1920s. The *Meditationes* also contained the first publication of Goldbach's conjecture (that every even integer greater than 2 can be written as the sum of two primes) and of Wilson's theorem (if p is a prime number, then $(p-1)! + 1$ is divisible by p), which was subsequently proved by LAGRANGE [VI.22].

Waring's problem and Goldbach's conjecture are discussed in PROBLEMS AND RESULTS IN ADDITIVE NUMBER THEORY [V.30].

VI.22 Joseph Louis Lagrange

In 1766 Lagrange left his native Turin, where he had been a founding member of what would later become the Turin Academy of Sciences, to become the mathematics director at the Berlin Academy of Sciences. In 1787 he moved to Paris to take up a position as a *pensionnaire veteran* at the Academy of Sciences. In Paris he also lectured at the École Polytechnique, founded in 1794, and served as one of the members of the committee that established the modern metric system.

Lagrange was only nineteen years old when he wrote to EULER [VI.19] announcing a new formalism to simplify Euler's method for finding a curve that satisfied an extremum condition. Lagrange's method was based on the introduction of a new differential operator, δ , to express the independent variations of the coordinates of a curve that produced a local infinitesimal deformation.

Using this formalism he derived the differential equation known today as the Euler-Lagrange equation, the fundamental equation of the CALCULUS OF VARIATIONS [III.96]. Suppose that we wish to find the function $y = y(x)$ that maximizes or minimizes a definite integral of the form

$$\int_a^b f(x, y, y') dx,$$

where $y' = dy/dx$. The equation states a necessary condition that this function must satisfy:

$$\frac{\partial f}{\partial y} - \frac{d}{dx} \left(\frac{\partial f}{\partial y'} \right) = 0.$$

This is a typical example of Lagrange's reductionist style. Throughout his career he sought suitable formalisms with which to express and solve the key problems of mathematical analysis.

Lagrange publicly presented his δ -formalism in a memoir published in the second volume (1760–61) of *Miscellanea Taurinensia*, a review he had helped to found. He coupled this memoir with another in which he used the same formalism to formulate a generalized version of the principle of least action (previously introduced by Maupertuis and Euler). As a result, he was able to derive the equations of motion of any system of distinct bodies attracted by central forces depending on the distances from their centers.

Meanwhile, in the first volume of *Miscellanea Taurinensia* (1759), Lagrange had published a memoir presenting a new approach to the problem of the vibrating string, where the string is first represented as a discrete system of n particles, and n is then allowed to tend to infinity. Using this method Lagrange argued that Euler was right to allow a large class of “functions,” both “continuous” and “discontinuous,” as solutions of this problem, whereas D’ALEMBERT [VI.20] had maintained that only “continuous functions” (that is, curves expressed by a single equation) could count as solutions.

Lagrange established a very general program in the foundations of classical mechanics in these memoirs. It was based on the interpretation of a continuous system as a limiting case of a discrete one and the use of the method of indeterminate coefficients: supposing that $P(x)$ is a polynomial in x , whose coefficients a_i ($i = 0, \dots, n$) depend on some indeterminates, and that $P(x) = 0$ for all x (in a given interval), this method consists of deducing the system of equations $\{a_i = 0\}_{i=0}^{i=n}$, from which the indeterminates are possibly determined. Lagrange extended this method to sums of polynomials in several (independent) variables, and (following Euler, d’Alembert, and many others) also used it with respect to power series. This program was further elaborated in two memoirs on the motion of the Moon (1764, 1780), and later realized in the *Mécanique Analytique* (1788). Here, the principle of least action was replaced by a generalization of the Bernoullian principle of “virtual velocities,” which were expressed by variations. Using what are now known as generalized coordinates, φ_i (that is, mutually independent coordinates in the configuration space of a discrete system that characterize completely the position of its bodies), Lagrange derived the equations that are

now named after him:

$$\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{\varphi}_i} \right) - \frac{\partial T}{\partial \varphi_i} + \frac{\partial U}{\partial \varphi_i} = 0,$$

where T and U are the kinetic and potential energy of the system, respectively.

The *Mécanique Analytique* appeared a century after NEWTON’s [VI.14] *Principia* and marked the culmination of a purely analytical approach to mechanics. In the preface Lagrange proudly stated that no diagrams would be found in the work and that everything would be reduced to “algebraic operations submitted to a regular and uniform progression.”

Lagrange made fundamental contributions to perturbation theory and THE THREE-BODY PROBLEM [V.36] with research published in the 1770s and 1780s. His methods were further developed by LAPLACE [VI.23] in his *Mécanique Céleste* and formed the basis for subsequent mathematical work in physical astronomy.

The method of indeterminate coefficients, or rather its extension to power series, is also the crucial technique underlying Lagrange’s approach to the calculus. In a memoir which appeared in the *Proceedings of the Berlin Academy* (1768) he used it to prove an important result connecting the calculus to the theory of algebraic equations, the so-called Lagrange inversion theorem, which states that a function $\psi(p)$ of a root p of an equation $t - x + \varphi(x) = 0$, where $\varphi(x)$ is an arbitrary function of x , can be expanded in a series based on the Taylor expansions of $\varphi(t)$ and $\psi(t)$ (the precise conditions to be satisfied by x , $\varphi(x)$, and $\psi(t)$ were later clarified by CAUCHY [VI.29] and ROCHÉ).

In a memoir of 1772 Lagrange returned to power series and proved that if a function $f(x + h)$ has a power-series expansion in h , then this series can be written in the form

$$\sum_{i=0}^{\infty} f^{(i)}(x) \frac{h^i}{i!},$$

where, for any i , $f^{(i+1)}$ is derived from $f^{(i)}$ in the same way that f' derives from f . Thus, he just had to prove, through an infinitesimal argument, that $f' = df/dx$ to conclude that the only power-series expansion of a function is its Taylor series. In *Théorie des Fonctions Analytiques* (1797) he then showed (or, rather, claimed to have shown), without making an appeal to the differential calculus, that any function $f(x + h)$ can be expanded in a power series, and suggested interpreting the differential formalism as a formalism that applies to the coefficients of $h^i/i!$ in such an expansion. In other words, he suggested defining the differential

ratios of any order (that is, the ratios $d^i y/dx^i$, where $y = f(x)$) as derivative functions supplying these coefficients, whereas previously these had been thought of as genuine ratios of differentials. He also proved that the remainder of a Taylor series could be written in the form now known as the Lagrange remainder.

The main results obtained by Lagrange within the theory of algebraic equations were presented in a long memoir of 1770 and 1771 in which the formulas for solving equations of degree 2, 3, and 4 were obtained through an analysis of permutations of the roots. This work constituted the starting point for the later researches of ABEL [VI.33] and GALOIS [VI.41]. In the same memoir Lagrange stated a particular but significant case of the theorem in group theory that today bears his name: that the order of a subgroup of a finite group divides the order of the group.

Lagrange also obtained important results in number theory. Arguably the most significant were the proof of a conjecture, advanced by FERMAT [VI.12] (among others) and which Euler had already tried to prove, asserting that any positive integer is the sum of (at most) four squares (1770), and the proof of Wilson's theorem (first guessed by Wilson and published by WARING [VI.21] without proof), asserting that if n is prime, then $(n-1)! + 1$ is divisible by n (1771).

Further Reading

Burzio, F. 1942. *Lagrange*. Torino: UTET.

Marco Panza

VI.23 Pierre-Simon Laplace

b. Beaumont-en-Auge, France, 1749; d. Paris, 1827
Celestial mechanics; probability; mathematical physics

Laplace is known to later mathematicians for many concepts of fundamental importance to mathematics: the LAPLACE TRANSFORM [III.93], the Laplace expansion, Laplace's angles, Laplace's theorem, Laplace functions, inverse probability, GENERATING FUNCTIONS [IV.18 §§2.4 §3], a derivation of the Gauss/Legendre least-squares rule of error by means of a linear regression, and the LAPLACIAN [I.3 §5.4] or potential function. Laplace developed the fields of celestial mechanics (the phrase was his coinage) and probability, and along the way the mathematics to service and advance them. For Laplace, celestial mechanics and probability were complementary instruments that implemented a unified

vision of a fully determined universe. Celestial mechanics vindicated the Newtonian system of the world. Probability was the measure, not of the operations of chance in nature, for there are none, but of human ignorance of causes, which was to be reduced to virtual certainty by calculation. The third reason for Laplace's importance to the history of science was the mathematization of physics in the first two decades of the nineteenth century. Apart from a few formulations—speed of sound, capillary action, refractive indices of gases—his role there was that of instigator and patron rather than of major contributor.

Laplace came up with the majority of the above concepts in a probabilistic context. The earliest hints of the method of solving difference, differential, and integral equations, later known as the Laplace transform, appeared in "Mémoire sur les suites" (1782a), where Laplace introduced generating functions. Laplace considered generating functions to be the approach of choice in solving problems that involved the development of functions in series and evaluation of the sums. Years later, in composing *Théorie Analytique des Probabilités* (1813), he subordinated all the analytical part to the theory of generating functions and treated the entire subject as their field of application. In the early memoir, however, he emphasized what he expected to be their applicability to problems of nature.

In an even earlier paper, "Mémoire sur la probabilité des causes par les événements" (1774), Laplace stated the theorem permitting the analysis later termed BAYESIAN [III.3]. Unknown to Laplace, Thomas Bayes had arrived at the same theorem eleven years earlier but had not developed it. Laplace for his part proceeded, in further investigations over some thirty years, to develop inverse probability into the basis for statistical inference, philosophical causality, estimation of scientific error, quantification of the credibility of evidence, and optimal voting rules in the proceedings of legislative bodies and judicial panels. His initial attraction to the approach was its applicability to human concerns. It was in the course of these papers, most notably "Mémoire sur les probabilités" (1780), that the word probability came to connote not merely the basic quantity in the theory of games and chance but a subject in itself.

Laplace first addressed error theory in the above paper on causality. The problem was to estimate the most appropriate mean value to be taken in a series of astronomical observations of the same phenomenon. He also determined how the limits of the error were

related to the number of observations (“Mémoire sur les probabilités” above). In “Essai pour connaître la population du Royaume” (1783–91) Laplace turned to demographic applications. In the absence of census data, one needed to determine the multiplier to be applied to the number of births at any one time in order to estimate the approximate size of a population. The specific problem Laplace solved was the size of the sample required for the probability of error to fall within given limits.

Laplace then put probabilistic investigation aside. Only twenty-five years later did he return to the subject, in the course of preparing the comprehensive *Théorie Analytique des Probabilités*. In 1810 he returned to the problem of determining the mean value from a large number of observations, which he interpreted as the problem of the probability that the mean value falls within certain limits. He proved a law of large numbers, stating that if positive and negative errors in an indefinitely large number of observations are assumed to be equally possible, then their mean result converges to a limit in a precise way. From this analysis followed the least-squares law of error. A priority dispute over the discovery of that law was even then simmering between GAUSS [VI.26] and LEGENDRE [VI.24].

In a long series of investigations brought together in *Traité de Mécanique Céleste* (1799–1825, five volumes), the two part “Mémoire sur la Théorie de Jupiter et de Saturne” (1788) demonstrates the most famous of his findings in planetary astronomy. He established that the current acceleration in orbital motion of Jupiter and the deceleration of Saturn are the reciprocal effects of their mutual gravitation, which are cyclical over many hundreds of years and not cumulative. From this and analysis of other phenomena that Laplace explored, it followed that the so-called secular inequalities of planetary motions are periodic over many centuries. Thus, they are not derogations from the law of gravity but evidence that its validity extended beyond the Sun-planet attractions that had been studied by NEWTON [VI.14]. He was never able to prove, however, that lunar acceleration is self-correcting over time.

The expansion known by Laplace’s name in the theory of determinants first appears in the background of the Jupiter-Saturn memoir in an analysis of the eccentricities and inclinations of orbits, “Recherches sur le calcul intégral et sur le système du monde” (1776). Except for that, Laplace’s mathematical originality is less notable in his analysis of planetary motion than in his development of the theory of probability.

More in evidence in his astronomical work is his motivational drive and his power and virtuosity in calculation, which may indeed have been more important throughout his long career. Laplace was masterful in finding rapidly convergent series, in obtaining mathematical expressions incorporating terms to represent a multitude of physical phenomena, in justifying the neglect of inconvenient quantities in order to reach solutions, and in giving the widest possible generality to his conclusions.

The attraction exerted by a spheroid on an external or internal point proved to be mathematically the most fertile set of problems in Laplacian planetary astronomy. In “Théorie des attractions des sphéroïdes et de la figure des planètes” (1785) Laplace employed LEGENDRE POLYNOMIALS [III.87] in a form later called Laplace functions. He also proved a theorem that stated that all ellipsoids with the same foci for their principal sections attract a given point with a force proportional to their masses. Laplace’s angles appear in his development of the equation for the attraction of a spheroid on a given point. Laplace used polar coordinates in this analysis. He transformed the equation into one in Cartesian coordinates in “Mémoire sur la théorie de l’anneau de Saturne” (1789). In 1828 George Green dubbed Poisson’s application of the formula to electrostatic and magnetic forces the potential function, the term used thereafter in classical physics.

Further Reading

The memoirs cited in this article can be found in the bibliography of C. C. Gillispie’s *Pierre-Simon Laplace: A Life in Exact Science* (Princeton University Press, Princeton, NJ, 1997).

For the mathematical content of Laplacian physics, see pp. 440–55 (and elsewhere) of I. Grattan-Guinness’s *Convolutions in French Mathematics* (Birkhäuser, Basel, 1990 (three volumes)).

Charles C. Gillispie

VI.24 Adrien-Marie Legendre

b. Paris, 1752; d. Paris, 1833

Analysis; theory of attractions; geometry; number theory

Legendre passed his career in Paris and seems to have been largely of independent means. Somewhat younger than LAGRANGE [VI.22] (who was resident there from 1787) and LAPLACE [VI.23], he did not quite match their

PUP: again you suggested square brackets here, and that would be fine by me. Before I make a global change throughout the Companion, though, I just want to confirm that you still want this change. (I think it was expressed as a preference rather than an instruction before.)

reputation, though the range of his mathematical interests was comparably wide. His professional appointments were modest; however, in 1799 he took over from Laplace as a graduation examiner at the École Polytechnique and remained there until his retirement in 1816. Additionally, in 1813 he succeeded Lagrange at the Bureau des Longitudes.

Legendre's early research concerned the shape of Earth and its external attraction to a point. Solutions of the differential equations involved led him to examine properties of the functions that are named after him; he was in rivalry with Laplace, after whom the functions were named during the nineteenth century. His other main concern in analysis, and the longest lasting, was with elliptic integrals. He wrote on them at great length up to a *Traité* of 1825–28. But in supplements of 1829–32 he acknowledged that his theory had just been eclipsed by the inverse ELLIPTIC FUNCTIONS [V.34] of JACOBI [VI.35] and ABEL [VI.33]. He also studied various other (functions defined as) integrals, including the beta and GAMMA FUNCTIONS [III.31]; solutions to differential equations; and optimization in the CALCULUS OF VARIATIONS [III.96].

Among Legendre's contributions to numerical mathematics was a beautiful theorem (found in 1789) relating spheroidal triangles (that is, triangles drawn on the surface of a spheroid) to spherical triangles, which was used in the 1790s by J.B.J. Delambre in the triangulation analysis that led to the specification of the meter. His most famous numerical result is the least-squares criterion of curve fitting, proposed in 1805 in connection with determining the orbits of comets. For him the criterion was simply one of minimization; he did not make the connections to probability theory that were soon to be effected by LAPLACE [VI.23] and GAUSS [VI.26].

Legendre's *Essai sur la Théorie des Nombres* (1798) was the first monograph on this subject. After reviewing CONTINUED FRACTIONS [III.22] and the theory of equations, he focused upon the algebraic branch, solving various Diophantine equations. Among many properties of integers, he stressed QUADRATIC RECIPROCITY [V.31], and proved various partition theorems concerning quadratic and some higher forms. Little in the book was new; while expanded editions appeared in 1808 and 1830, he had been quickly eclipsed on methods of proof by the *Disquisitiones Arithmeticae* (1801) of the young Gauss.

For educational use Legendre produced *Elements de Géométrie* (1794), an account of EUCLIDEAN GEOMETRY

[I.3 §6.2] that emulated the same kind of form and organization and standards of proof of the Greek original. He also handled aspects that had lain outside Euclid's concerns, such as alternatives to the parallel postulate, related numerical issues such as approximations to the value of π , and a lengthy summary of planar and spherical trigonometry. He produced eleven further editions up to 1823 and there were further posthumous editions up until 1839 (which were followed by reprints). It was a very influential book in mathematics education.

Further Reading

de Beaumont, E. 1867. *Eloge Historique de Adrien Marie Legendre*. Paris: Gauthier-Villars.

Ivor Grattan-Guinness

VI.25 Jean-Baptiste Joseph Fourier

b. Auxerre, France, 1768; d. Paris, 1830

Analysis; equations; heat theory

Unusually for a mathematician, Fourier pursued a distinguished nonmathematical career. He was a civilian member of General Bonaparte's expedition to Egypt (1798–1801), important enough for the First Consul to make him, in 1802, the Prefect of the *département* at Grenoble, a position which he held until Emperor Napoleon's fall in the mid 1810s. Thereafter, Fourier moved to Paris, where he managed to establish himself to the extent of being appointed a *secrétaire perpétuel* of the Paris Academy of Sciences in 1822.

The prefectureship involved heavy commitments, and Fourier was also active in Egyptology, most notably discovering a teenager named Jean Champollion in Grenoble, who was later to decipher the Rosetta Stone and who helped to found the discipline. Nevertheless, between 1804 and 1815 he also created most of his scientific work. His motivation was the mathematical study of the diffusion of heat in continuous and solid bodies; his “diffusion equation” for this purpose was not only novel in itself but also marked the first large-scale mathematization of physical phenomena that lay outside mechanics. To solve this differential equation he proposed using infinite trigonometric series. These series were already known but had a low status. Fourier (re)found many properties: not only the formulas for their coefficients and some conditions for their convergence but especially their representability, namely, how a periodic series could represent a general function. For

PUP: this suggested as alternative to what came here before. OK?

diffusion in a cylinder he found many properties of the Bessel function $J_0(x)$, which was then little studied.

Fourier presented his findings to the scientific class of the Institut de France in 1807. LAGRANGE [VI.22] did not like the series, while LAPLACE [VI.23] was disappointed in the physical modeling. But Laplace also gave him a clue about solutions of the diffusion equation for *infinite* bodies, which led Fourier to find, by 1811, his integral solution (including inversion) for them. His main publication was the book *Théorie Analytique de la Chaleur* (1822), which greatly influenced younger mathematicians: for example, the first satisfactory proof of the convergence of the series by DIRICHLET [VI.36] (1829) and their use in fluid dynamics by C.L.M.H. Navier (1825). Less happy was his relationship with POISSON [VI.27], who tried to rederive the entire theory following the molecularist physical principles of Laplace and the methods of solution of Lagrange, but only added a few special cases.

Fourier also worked on other topics in mathematics. As a teenager he gave the first proof of DESCARTES's [VI.11] rule of signs on the numbers of positive and negative roots of a polynomial equation. (He used an inductive proof that has now become standard.) He also found an upper bound on the number of roots within a given interval, which J.C.F. Sturm improved to an exact evaluation in 1829. At that time Fourier was trying to finish a book on equations, which appeared posthumously in 1831 thanks to Navier. The main novelty was the basic theory of LINEAR PROGRAMMING [III.86], as we now call it. Despite his prestige and advocacy, he gained few followers (Navier was one), and the theory lay dormant for over a century. Fourier also took up a few aspects of Laplace's work on mathematical statistics, examining the status of the NORMAL DISTRIBUTION [III.73 §5].

Further Reading

Fourier, J. 1888–90. *Oeuvres Complètes*, edited by G. Darboux, two volumes. Paris: Gauthier-Villars.

Grattan-Guinness, I., and J. R. Ravetz. 1972. *Joseph Fourier*. Cambridge, MA: MIT Press.

Ivor Grattan-Guinness



b. Brunswick, Germany, 1777; d. Göttingen, Germany, 1855
Algebra; astronomy; complex function theory including elliptic function theory; differential equations; differential geometry; land surveying; number theory; potential theory; statistics

Gauss's prodigious mathematical abilities brought him to the attention of the duke of Brunswick when he was fifteen, when the duke paid for his further education, lifting him out of near poverty. For the rest of his life Gauss felt a loyalty to the state and a strong desire to do useful work, which led him to become a professional astronomer. In 1801 he was the first person to manage to reobserve Ceres, the first asteroid to be discovered, after it had disappeared behind the Sun. Gauss produced a novel statistical analysis of the original observations, using the method of least squares, which he had invented but not published, to predict where Ceres would reappear. Gauss then assisted for many years in the analysis of the orbits of several more asteroids. He also wrote extensively on celestial mechanics and cartography, and did important work on telegraphy.

Nonetheless, it is as a pure mathematician that Gauss will always be remembered. In 1801 he published his *Disquisitiones Arithmeticae*, the book that created modern algebraic number theory. In it he gave the first rigorous proof of the law of QUADRATIC RECIPROCITY [V.31], going on to find five more proofs over the years. Later he extended the theorem to higher powers, introducing the Gaussian integers for the purpose in 1831 (Gaussian integers are numbers of the form $m + ni$, where m and n are integers and $i = \sqrt{-1}$). He did major work on differential equations, chiefly the hypergeometric equation, which is a second-order lin-

ear differential equation depending on three parameters and having two singular points, with the property that many of the familiar functions of analysis are related to its solutions. He showed that this equation played a significant role in the new theory of ELLIPTIC FUNCTIONS [V.34], but because most of this work was unpublished it had no influence on the dramatic and rapidly advancing publications of ABEL [VI.33] and JACOBI [VI.35]. This unpublished work showed that he was the first mathematician to see the need to create a theory of complex functions of a complex variable. He also gave four proofs of THE FUNDAMENTAL THEOREM OF ALGEBRA [V.15]. By the 1820s he was persuaded that physical space might not be Euclidean, but he confined his opinion to his circle of friends, most of them astronomers and sympathetic to the idea; the much more detailed accounts of BOLYAI [VI.34] and LOBACHEVSKII [VI.31] were published independently in the early 1830s. Credit for the first detailed, mathematical descriptions of a non-Euclidean space therefore rightly attaches to Bolyai and Lobachevskii (for further discussion of this, see GEOMETRY [III.2 §7]). In 1827 Gauss wrote his *Disquisitiones Generales Circa Superficies Curvas*, in which the concept of intrinsic (Gaussian) curvature of a surface was put forward for the first time, thus reformulating differential geometry.

In statistics, he was one of the two or three discoverers of the NORMAL DISTRIBUTION [III.73 §5], and he was an expert in error analysis, bringing the levels of accuracy in astronomy to land surveying. In that context he invented the heliotrope, which couples a mirror to a telescope in order to transmit a precise beam of light, to improve precision measurement.

The sheer volume of Gauss's work is overwhelming. The *Werke* run to twelve volumes, and there are several books, of which the *Disquisitiones Arithmeticae* stands out.

A truly original mathematician and scientist, Gauss was otherwise a conservative in his tastes and views. His first marriage ended after only four years with the death of his wife in 1809; he then married again. A number of Gauss's descendants may now be found in the United States.

Gauss was the last great mathematician to be called the "Prince of mathematicians," and he has been admired as much for his breadth as for the depth of his insights and the fertility of his ideas. His own view of mathematics and its importance is captured both in the much-quoted remark that "mathematics is the queen of the sciences and arithmetic the queen of mathematics"

(which he did say) and in the apocryphal remark that "mathematics is the queen and the servant of science."

Further Reading

Dunnington, G. W. 2003. *Gauss: Titan of Science*, new edition with additional material by J. J. Gray. Washington, DC: Mathematical Association of America.

Jeremy Gray

VI.27 Siméon-Denis Poisson

b. Pithiviers, France, 1781; d. Paris, 1842

Analysis; mechanics; mathematical physics; probability

A brilliant graduate of the École Polytechnique in 1800, Poisson was quickly appointed to the staff, and became professor and graduation examiner there until his death. He was also founder professor of mechanics at the new Paris Faculté des Sciences of the Université de France; from 1830 he was also a member of the governing Council of the Université.

Poisson's research output was dominated by his adherence to the traditions established by LAGRANGE [VI.22] and LAPLACE [VI.23]. Like Lagrange he preferred to algebraize theories, and to rely if possible upon power series and variational methods. From the mid 1810s he challenged the new theories of FOURIER [VI.25] (especially the solving of differential equations by trigonometric series and the Fourier integral) and of CAUCHY [VI.29] (the new approach to real-variable analysis using limits, and his innovation of complex-variable analysis). His overall achievements were much less significant than theirs: the main novelties were the "Poisson integral," which embedded Fourier series within a power series; and a summation formula. He also studied the general and singular solutions of differential, difference, and mixed equations.

In physics Poisson tried to justify Laplace's claim that all physical phenomena were molecular, and that the cumulative action upon a molecule of all its companions should be expressed mathematically in terms of an integral. He applied this approach to heat diffusion and to elasticity theory by the mid 1820s, but then decided that integrals should be replaced by sums; he elaborated this alternative especially in capillary theory (1831). Curiously, molecularism did not dominate his most important contributions to physics: to electrostatics (1812–14) and to magnetic bodies and the process of magnetization (1824–27). His mathematical contributions to these topics included modifying

Laplace's equation to what we now call Poisson's equation, which deals with the potential at points inside a charged body or region of charge (1814); and also a divergence theorem (1826).

In mechanics, between 1808 and 1810 Poisson and Lagrange developed the brackets theory (named after them) of canonical solutions to the equations of motion. Poisson's motivation was to extend, to second-order terms in masses of the planets, Lagrange's superb attempt to prove that the planetary system was stable; in later work he examined this (first-order) problem specifically, as well as other aspects of perturbation theory. He also analyzed rotating bodies by using moving frames of reference (1839), in an analysis that was to inspire Léon Foucault to propose his famous long pendulum in 1851. His best-known publications include a substantial and wide-ranging two-volume *Traité de Mécanique* (editions in 1811 and 1833), which did not, however, have room for Louis Poincaré's beautiful recent theory (1803) of the couple in statics. In the mid 1810s, he studied deep-body fluid dynamics, in rivalry with Cauchy.

Poisson was one of the few contemporaries to take up Laplace's work in probability theory and mathematical statistics. He studied various PROBABILITY DISTRIBUTIONS [III.73]: not only the one named after him (1837, rather in passing) but also the so-called Cauchy (1824) and Rayleigh (1830) distributions. He also examined proofs of the CENTRAL LIMIT THEOREM [III.73 §5], and formulated THE LAW OF LARGE NUMBERS [III.73 §4] (his term). One of his main applications was to the old problem of determining the probability that a triad of judges would come to the correct decision in court cases (1837).

Further Reading

Grattan-Guinness, I. 1990. *Convolutions in French Mathematics 1800–1840*. Basel: Birkhäuser.

Métivier, M., P. Costabel, and P. Dugac, eds. 1981. *Siméon-Denis Poisson et la Science de son Temps*. Paris: École Polytechnique.

Ivor Grattan-Guinness

VI.28 Bernard Bolzano

b. Prague, 1781; d. Prague, 1848

Catholic priest and Professor of Theology, Prague (1805–19)

Bolzano was concerned with problems connected with finding the “correct,” or the most appropriate, proofs and definitions in analysis and related areas. In 1817 he proved an early version of the intermediate-value theorem for continuous functions—he was among the first to have a rigorous conception of a continuous function—and in the course of doing so proved the following important lemma. *If a property M does not apply to all values of a variable x but does apply for all values smaller than a certain u , then there is always a quantity U , which is the greatest of those of which it can be asserted that all smaller x possess the property M .* The value u in this formulation is a lower bound for the (nonempty) set of numbers with the property not- M . Bolzano's lemma is therefore equivalent to what nowadays might be called the “greatest lower bound” axiom (or, more commonly and equivalently, the “least upper bound” axiom). It is also equivalent to the Bolzano-Weierstrass theorem (that every bounded infinite set in \mathbb{R} , or more generally \mathbb{R}^n , has an accumulation point). It is likely that WEIERSTRASS [VI.44] independently rediscovered the Bolzano-Weierstrass theorem, but it is also likely that he knew, and was influenced by, Bolzano's proof technique of iterated bisection (used by Bolzano in 1817).

In the early 1830s it was widely believed that a continuous function must be differentiable except at some isolated points. But at that time Bolzano constructed a counterexample (although he did not publish it), and proved that it was such—more than thirty years before the well-known counterexample due to Weierstrass.

Bolzano had a surprising variety of insights and successful proof techniques that were well ahead of their time: notably in analysis, topology, dimension theory, and set theory.

VI.29 Augustin-Louis Cauchy

b. Paris, 1789; d. Sceaux, France, 1857

Real and complex analysis; mechanics; number theory; equations and algebra

Trained as a roads and bridges engineer at the École Polytechnique (hereafter, “EP”) and the École des Ponts et Chaussées (1805–10), Cauchy passed his career as an academic at the EP and the Paris Faculté des Sciences of the Université de France until 1830, when he left France with the deposed royal family after the revolution of that year. He returned only in 1838, and later taught in the Paris Faculté.

Of Cauchy's many contributions to pure and applied mathematics, the best known are in mathematical

analysis. In the foundations of real variables, he replaced all previous approaches to the theory with one that (in more developed forms) has now become standard: (i) lay down an explicit *theory* of limits; (ii) formulate definitions carefully, and in general terms; (iii) define the derivative of a function as the limiting value of the difference quotient, its integral as the limiting value of a sequence of partition sums, its continuity in terms of the joint passage to limits of any sequence of its argument and of its corresponding values, and the sum of a convergent infinite series as the limiting value of its partial sums. A key ingredient in all this was the idea that (iv) limits may not exist: their existence has to be justified carefully. Similarly, (v) the existence of solutions to differential equations has to be proved, not just assumed.

This approach brought a new level of rigor to analysis; for example, for the first time THE FUNDAMENTAL THEOREM OF CALCULUS [I.3 §5.5] was a genuine theorem, governed by conditions on the function. However, this emphasis on limits made the theory hard for beginners: it was not liked by staff or students at the EP, where he taught it in this form between 1816 and 1830 and published it extensively, especially in his *Cours d'Analyse* (1821) and his *Résumé* (1823) of the calculus. Its rise to standard educational practice was very gradual, both in France and elsewhere.

Another major innovation of Cauchy dates from 1814, when he began to create complex-variable analysis. Initially the integrand was a complex function but the limits of integration were real; however, from 1825 on they too became complex, and in this form he found many theorems on the residues of functions over closed domains of various shapes. Unusually for him, his progress was fitful, and he cast the theory in terms of the complex plane only in the mid 1840s. He also studied the general theory of complex functions, including their expansion in power series of various kinds.

Cauchy's main single achievement in applied mathematics lies in linear elasticity theory, where in the 1820s he used stress-strain models to analyze the behavior of various kinds of surfaces and solids; later he adapted it to study aspects of (aetherian) optics. In the 1810s he studied deep-body fluid dynamics, where he found Fourier-integral solutions. In this and several other areas he was in some competition with Fourier and, especially, Poisson, regarding both the quality of the theory and the chronology of its development.

Cauchy's other contributions lie in basic mechanics (derived from the EP teaching); singular and general solutions of differential equations; the theory of equations, especially methods that helped in the rise of group theory; algebraic number theory; perturbation theory in celestial mechanics; and an astounding paper of 1829 on quadratic forms, which could have launched the spectral theory of matrices had its author recognized its significance!

Further Reading

- Belhoste, B. 1991. *Augustin-Louis Cauchy. A Biography*. New York: Springer.
- Cauchy, A. L. 1882–1974. *Oeuvres Complètes*, twelve volumes in the first series and fifteen in the second. Paris: Gauthier-Villars.

Ivor Grattan-Guinness

VI.30 August Ferdinand Möbius

b. Schulpforta, Saxony, 1790; d. Leipzig, Germany, 1868
Astronomy; geometry; statics

Möbius was briefly a student of GAUSS [VI.26], and worked as an astronomer at Leipzig University for almost all of his life. His finest mathematical work was his *Der barycentrische Calcul* (1829), in which he introduced algebraic methods into the study of projective geometry. He showed in this way how points can be described by a homogeneous triple of coordinates, lines can be described by linear equations, the concept of cross-ratio can be introduced, and the duality of points and lines in the plane can be handled algebraically. He also introduced a *Möbius net*, which is the projective equivalent of squared paper in Cartesian geometry. His work is all the more remarkable because Möbius knew very little of Poncelet's radical reinvention of projective geometry only a few years before. In its turn his work was for a time overshadowed by Jakob Steiner's synthetic treatment of projective geometry of 1832, and then Plücker's two books on algebraic curves in the 1830s, but the simplicity and generality of Möbius's methods were important in establishing projective geometry as a rigorous mainstream subject.

In the 1830s Möbius developed a geometrical theory of statics and the composition of forces, and it was in this connection that he showed that whereas duality in plane geometry necessarily gives rise to a conic, duality in space need not. Möbius's study of duality in space, which pairs points with planes, led him to consider the

set of all lines in space, which is a four-dimensional space. It pleased the educator Rudolf Steiner very much that the ordinary three-dimensional space may also be thought of as a four-dimensional space, because Steiner's philosophy was directed against breaking what he saw as a stranglehold of orthodox teaching.

Möbius is also remembered for the "Möbius band" (or MÖBIUS STRIP [IV.7 §2.3]), a one-sided or nonorientable surface, but the first mathematician to describe such a surface was his compatriot J. B. Listing, in July 1858 (published in 1861). Möbius discovered it only in September 1858 (publishing it in 1865). He is also one of the most important mathematicians to study inversion in circles, and his account of it in 1855 is one reason that such transformations are often called Möbius transformations.

Further Reading

Fauvel, J., R. Flood, and R. J. Wilson, eds. 1993. *Möbius and His Band*. Oxford: Oxford University Press.

Möbius, A. 1885–87. *Gesammelte Werke*, edited by R. Baltzer (except volume 4, edited by W. Scheibner and F. Klein), four volumes. Leipzig: Hirzel.

Jeremy Gray

VI.31 Nicolai Ivanovich Lobachevskii

b. Nizhni Novgorod (formerly Gorki), Russia, 1792;

d. Kazan, Russia, 1856

Non-Euclidean geometry

Lobachevskii came from a poor background, but his mother was able to have him enrolled at the local Gymnasium (or high school) on a scholarship in 1800. In 1805 the Gymnasium was made the kernel of the new University of Kazan, and in 1807 Lobachevskii began to study there. The university had just appointed Martin Bartels as Professor of Mathematics, and Bartels not only trained Lobachevskii well, but protected him from trouble with the authorities when Lobachevskii was suspected of atheism. Eventually, Lobachevskii graduated not with the ordinary degree but with a Master's qualification, and his career as a professional mathematician began.

In 1826, after a reform of the university, Lobachevskii gave a public lecture: "On the principles of geometry, with a rigorous demonstration of the theory of parallels." The manuscript of this talk is now lost, but it probably marked the start of Lobachevskii's awareness of a non-Euclidean geometry. Lobachevskii was

soon elected Rector of the University of Kazan, a post he occupied with distinction for thirty years, helping to protect the university from a cholera epidemic in 1830, to rebuild it after a fire in 1841, and generally to expand its library and other facilities.

In the 1830s he also wrote his major works, on a geometry different in only one respect from Euclidean geometry. He called it imaginary geometry and it is known today as non-Euclidean geometry. In the new geometry, given a line in a plane and a point not on the line there are two lines through the point that are asymptotic to the given line (one in each direction); these two lines separate the lines through the point which meet the given line from those that do not. Lobachevskii called these the *parallels* to the given line through the given point. Starting from this definition, he gave formulas for the new trigonometry of triangles, and showed that these formulas reduce to the familiar formulas of plane Euclidean trigonometry when the triangles are very small. He extended his results to describe a geometry of three dimensions, thus making it clear that his new geometry could be a geometry of space, and attempted, inconclusively, to measure the parallax of stars in order to determine whether his imaginary geometry gave a more accurate account of space than Euclidean geometry.

He published these conclusions in lengthy papers in Russian in the *Journal of Kazan University*, but they drew only a relentlessly hostile review from Ostrogradskii, a much better known mathematician in Saint Petersburg. He published in French in a German journal in 1837, in German in a booklet of 1840, and again in French in 1855, but to little avail. GAUSS [VI.26] appreciated the booklet of 1840 and in 1842 had Lobachevskii made a corresponding member of the Göttingen Academy of Sciences, but this was to be the only acclaim Lobachevskii received in his lifetime.

Lobachevskii's final years were marked by terrible financial and mental decline. Such was the chaos of his household that Lobachevskii's biographers have been unable to establish the number of children born into it, but it may well have been fifteen or even eighteen.

Further Reading

Gray, J. J. 1989. *Ideas of Space: Euclidean, Non-Euclidean, and Relativistic*, second edn. Oxford: Oxford University Press.

Lobachetschefsckij, N. I. 1899. *Zwei geometrische Abhandlungen*, translated by F. Engel. Leipzig: Teubner.

Rosenfeld, B. A. 1987. *A History of Non-Euclidean Geometry: Evolution of the Concept of a Geometric Space*. New York: Springer.

Jeremy Gray

VI.32 George Green

b. Nottingham, England, 1793; d. Nottingham, 1841

Miller; *Fellow of Caius College, Cambridge (1839–41)*

Green, a self-taught mathematician, went to Cambridge University at the age of forty, having already produced his most important work, the privately printed *An Essay on the Application of Mathematical Analysis to the Theories of Electricity and Magnetism* (1828). In this work, which Green opened by stressing the central role of the “potential function” (a term that he himself coined), he proved the three-dimensional version of the theorem now named after him, and introduced the concept that RIEMANN [VI.49] later called Green’s function (1860). The *Essay* became widely known only after its discovery in 1845 by William Thomson (later Lord Kelvin), who was responsible for its republication in the *Journal für die reine und angewandte Mathematik* (1850–54).

Green gave his version of the theorem (in modern notation) as

$$\iiint U \Delta V \, dv + \iint U \frac{\partial V}{\partial n} \, d\sigma = \iiint V \Delta U \, dv + \iint V \frac{\partial U}{\partial n} \, d\sigma,$$

where U and V are two continuous functions of x, y, z whose derivatives are not infinite at any point of an arbitrary body, n is the surface normal of the body directed inward, and $d\sigma$ is a surface element. The result today known as Green’s theorem, which is the planar version of the above, was first published by CAUCHY [VI.29] in 1846, and it can be given (in modern notation) as follows. Let R be a closed plane region with a piecewise-smooth boundary curve C with positive orientation. Let $P(x, y)$ and $Q(x, y)$, having continuous partial derivatives, be defined on an open region containing R . We then have

$$\int_C (P \, dx + Q \, dy) = \iint_R \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dx \, dy.$$

More original than his theorem, however, was the powerful technique Green developed to solve certain second-order differential equations. In essence, Green sought a “potential function” and formulated the conditions it needed to satisfy. His great insight was to recognize that the central issue in potential theory was

to relate properties inside volumes to properties on their surfaces. Green’s functions are extensively used today in the solution of inhomogeneous differential equations with boundary conditions and in the solution of partial differential equations.

VI.33 Niels Henrik Abel

b. Finnøy, Norway, 1802; d. Froland, Norway, 1829

Theory of equations; analysis; elliptic functions; Abelian integrals

Abel’s life was short and penurious, but successful, and he received recognition in his lifetime. His father—a minister of the church in Norway, but also at one time a government minister—overreached himself and when he died he left the family in straitened circumstances. Abel’s exceptional intellectual talents were recognized at school, and funds were raised to enable him to complete his education and, in particular, to study mathematics. At age twenty-two, he was awarded a scholarship to make a two-year tour of Europe, during which he studied in Berlin and Paris. In Berlin he met and was befriended by Auguste Crelle, the engineer who had just founded the *Journal für die reine und angewandte Mathematik* (otherwise known as *Crelle’s Journal*). Almost all of Abel’s mathematical work was published in the first four volumes of the journal. From 1826 until his death in 1829 Abel eked out a poor existence, earning a little by teaching, but using what few resources he had to support his mother and his younger brother. He died of consumption at the age of twenty-seven within a couple of days of the news reaching Norway that he had been appointed to an established post in Berlin.

Abel’s main mathematical contributions lie in three distinct areas. The first of these was the theory of equations. Here he was influenced by ideas published by LAGRANGE [VI.22] in 1770 and CAUCHY [VI.29] in 1815 about the form of functions of the roots of an equation, and what happens to such functions when the roots are permuted. Lagrange had hinted at the possibility that quintic equations might perhaps not be soluble in classical terms and Paolo Ruffini had expended much effort between 1799 and 1814 trying to prove this, though he had not managed to persuade his contemporaries. Abel’s first success was to give an acceptable proof of the fact that, for polynomial equations of degree 5, there is no formula in the coefficients involving the usual operations of arithmetic together with extraction of roots that will always yield a solution. This first appeared in 1824 in a short pamphlet written

in French and published privately in Christiania (Oslo). Once Abel reached Berlin, however, Crelle translated it into German and published it in the first volume of his journal; he also published a fuller, more detailed account, covering polynomials of any degree greater than 4, in 1826.

Abel returned to equations a few years later, publishing a long paper in 1829 about a class of equations satisfying two special conditions. His first requirement is that every root of the equation can be expressed as a function of every other root, the second that these functions commute (in modern terms, the GALOIS GROUP [V.24] of the equation is commutative). He proved various theorems about such equations, the most striking being that they are soluble by radicals. This represented an extensive generalization of the ideas described by GAUSS [VI.26] in the seventh part of *Disquisitiones Arithmeticae*, where the special case of cyclotomic equations (which satisfy both of these conditions) is treated systematically. It was in honour of this work that, later, the adjective “Abelian” was applied to groups that are commutative. It is important to appreciate, however, that Abel reached his results in the theory of equations without any appeal to groups, which at that time were not yet known.

He also made major contributions to the theory of convergence. Although there had been over a century of critical thinking devoted to foundations of the calculus, modern ideas of rigour were only just emerging in the writings of BOLZANO [VI.28], Cauchy, and others. Convergence had received some attention in Cauchy’s lectures of 1820–21, but series in general, and power series in particular, were still far from well understood. Among other contributions, Abel offered a proper proof of the binomial theorem for exponents other than positive integers, and the insight about the continuity of a function defined by a power series as its argument goes to the circle of convergence that is now known as Abel’s limit theorem.

Perhaps his greatest discoveries, however, were in the area where analysis and algebraic geometry come together. To summarize his legacy in this area in just a few words: first, a new and productive approach to the theory of ELLIPTIC FUNCTIONS [V.34]; and, second, a vast generalization of elliptic functions to what are now called Abelian functions and Abelian integrals. In this area Abel competed for priority with JACOBI [VI.35]. Most (though by no means all) of his work was written in two memoirs. One was published in two parts, “Recherches sur les fonctions elliptiques” and “Précis

d’une théorie des fonctions elliptiques,” coming to well over two hundred pages in *Crelle’s Journal* in 1828 and 1829. The other, entitled “Mémoire sur une propriété générale d’une classe très étendue de fonctions transcendentes,” was submitted to the Paris Academy of Sciences in October 1826. There it lay on Cauchy’s desk, unread until after Abel’s death. It was published by the Paris Academy in 1841. The manuscript itself, however, was stolen by G. Libri, lost, and rediscovered in parts between 1952 and 2000 by Viggo Brun and Andrea del Centina.

In June 1830 the Paris Academy awarded its Grand Prix de Mathématiques jointly to Abel (posthumously) and Jacobi for their work on elliptic functions.

Further Reading

- Del Centina, A. 2006. Abel’s surviving manuscripts including one recently found in London. *Historia Mathematica* 33:224–33.
- Holmboe, B., ed. 1839. *Œuvres Complètes de Niels Henrik Abel*, two volumes. (Second edn.: 1881, edited by L. Sylow and S. Lie. Christiania: Grøndahl & Søn.)
- Ore, O. 1957. *Niels Henrik Abel: Mathematician Extraordinary*. Minneapolis, MN: University of Minnesota Press. (Reprinted, 1974. New York: Chelsea.)
- Stubhaug, A. 1996. *Et Foranskutt Lyn: Niels Henrik Abel Og Hans Tid*. Oslo: Aschehoug. (English translation: 2000, *Niels Henrik Abel and His Times: Called Too Soon by Flames Afar*, translated by R. H. Daly. New York: Springer.)

Peter M. Neumann

VI.34 János Bolyai

b. Klausenburg, Transylvania, Hungary (now Cluj, Romania), 1802;
d. Marosvásárhely, Hungary (now Tirgu-Mures, Romania), 1860
Non-Euclidean geometry

János Bolyai’s father Farkas Bolyai taught him mathematics at home, using the first six books of EUCLID’s [VI.2] *Elements* and EULER’s [VI.19] *Algebra*. From 1818 to 1823 János studied at the Royal Engineering Academy in Vienna, and then served as an engineer in the Austrian Army for ten years, before retiring on a pension as a semi-invalid. Probably inspired by his father’s attempts to prove the parallel postulate, a key assumption in Euclidean geometry, but very much against the advice of his father, János also attempted to prove it. But in 1820 he switched direction and attempted to show that there could be a geometry independent of the parallel postulate. By 1823 he believed he had succeeded, and after much subsequent

Jeremy Gray

discussion father and son agreed to publish the son's ideas as a twenty-eight-page appendix to his father's two-volume work on geometry in 1832.

In this appendix, Bolyai started from a new definition of parallels, according to which, given a line in a plane and a point not on the line, there are many lines through the point that do not meet the given line. Of these lines, there are two that are asymptotic to the given line (one in each direction), and Bolyai called these the parallels to the given line through the given point. He went on to derive many results that follow from this assumption in the geometry of two and three dimensions, and gave formulas for the new trigonometry of triangles. He showed that these formulas reduce to the familiar formulas of plane Euclidean geometry when the triangles are very small. He also found a surface in his three-dimensional geometry in which geometry is Euclidean. He concluded that there were logically two geometries and that it remained undecided which one corresponded to reality. He also showed that in his new geometry it was possible to construct a square equal in area to a given circle, thus accomplishing a feat that was widely (and, as was later shown, correctly) believed to be impossible in Euclidean geometry.

A copy of the book was sent to GAUSS [VI.26], who eventually replied on March 6, 1832, that he could not praise the work, for "to praise it, would be to praise myself," going on to claim that the methods and results in the appendix agreed with his own work over the previous thirty-five years, although he was "very glad that it was just the son of my old friend, who takes the precedence of me in such a remarkable manner." This endorsement of the validity of János's ideas pleased the father but infuriated the son, and soured relations between father and son for several years. They did eventually resume an uncomfortable relationship, which persisted until Farkas's death in 1856.

János Bolyai published virtually nothing else, and his discovery was not appreciated in his lifetime. Indeed, it is unclear that anyone but Gauss ever read it, but specific comments about it that Gauss left behind led mathematicians back to it, and it was translated into French by Hoüel in 1867 and into English in 1896 (reprinted in 1912 and 2004).

Further Reading

Gray, J. J. 2004. *János Bolyai, Non-Euclidean Geometry and the Nature of Space*. Cambridge, MA: Burndy Library, MIT Press.

VI.35 Carl Gustav Jacob Jacobi

b. Potsdam, Germany, 1804; d. Berlin, 1851

Theory of functions; number theory; algebra; differential equations; calculus of variations; analytical mechanics; perturbation theory; history of mathematics

Jacobi grew up as Jacques Simon Jacobi in a wealthy and well-educated Jewish family. He was baptized during his first year at the University of Berlin in 1821, probably in order to make it possible for him to follow an academic career at a time when Jews were ineligible for academic positions. Jacobi studied classics under the famous philologist Boeckh and philosophy under Hegel. Owing to the mediocrity of the mathematics staff in Berlin at that time, he was self-taught in the discipline, which soon became his favorite. He read EULER [VI.19], LAGRANGE [VI.22], LAPLACE [VI.23], GAUSS [VI.26], and, last but not least, Greek mathematicians like Pappus and Diophantus. In 1825, Jacobi was awarded his doctorate for a thesis, written in Latin, on the theory of functions. The subsequent *disputatio* (discussion) included critical comments both on Lagrange's theory of functions and on his analytical mechanics. The following year Jacobi went to the University of Königsberg, where (in 1829) he got a full professorship. In 1834, he and the physicist F. E. Neumann founded the "Königsberg mathematical physics seminar," which, because of the close connection between research and teaching that it fostered, soon led to Königsberg becoming the most successful and influential educational institution for theoretical physics and mathematics in the German-speaking part of the scientific world. By 1844, when Jacobi left Königsberg because of poor health and in order to become a member of the Berlin Academy of Sciences, he was recognized as Germany's most important mathematician after Gauss. After seven more fruitful years of research in Berlin he died unexpectedly from smallpox.

Throughout his life Jacobi was an advocate of pure mathematics, conceiving mathematical thinking as a means of developing the human intellect and, indeed, of advancing humanity itself. His first published paper (1827), which was influenced by Gauss's *Disquisitiones Arithmeticae*, was devoted to number theory (cubic residues). Further investigations were devoted to higher residues, the division of the circle, quadratic forms, and related subjects. Many of Jacobi's results in number theory were published in the book *Canon*

Arithmetica (1839). The extension of the concept of divisibility to algebraic numbers by Jacobi and Gauss paved the way for the later algebraic theory of numbers (by KUMMER [VI.40] and others).

Jacobi's "most original achievements" (in the words of KLEIN [VI.57]) were his contributions to the theory of ELLIPTIC FUNCTIONS [V.34], which developed in competition with ABEL [VI.33] between 1827 and 1829. Starting with LEGENDRE's [VI.24] work, Jacobi's approach was analytical and focused on the transformation of elliptic functions, their properties (like double periodicity), and the introduction of the inverse function. Jacobi's research on elliptic functions culminated in the book *Fundamenta Nova Theoriae Functionum Ellipticarum* (1829). Together with Abel he should be viewed as one of the founders of the theory of complex functions, which emerged in the second half of the century. In particular, his application of research on elliptic functions to Diophantine equations became important for the development of analytic number theory. Jacobi's contributions to algebra include investigations into the theory of determinants (the "Jacobian" functional determinant) and their relation to inverse functions, into quadratic forms ("Sylvester's law of inertia"), and into the transformation of multiple integrals.

Even Jacobi's work in mathematical physics bears the stamp of "pure mathematics": following the analytical tradition of Euler and Lagrange, he presented the foundations of mechanics in an abstract and formal manner, paying special attention to the relation between CONSERVATION LAWS [IV.12 §4.1] and symmetries of space and to the unifying role of variational principles. Jacobi's achievements in this area, which he developed in close relation to the theory of differential equations and the CALCULUS OF VARIATIONS [III.96], include what is now called the "Jacobi-Poisson theorem," the "principle of the last multiplier," a theory for integrating HAMILTON's [VI.37] CANONICAL EQUATIONS OF MOTION [IV.16 §2.1.3] by transformation ("Hamilton-Jacobi theory"), and a time-independent formulation of the principle of least action ("Jacobi's principle"). His approach to these areas and the results he obtained are documented in two comprehensive books based on his lectures: *Vorlesungen über Dynamik* (1866) and *Vorlesungen über Analytische Mechanik* (not published until 1996). The former had considerable impact on the development of German mathematical physics in the last third of the nineteenth century. The latter reveals Jacobi's criticism of the traditional understanding of mechanical principles (as laws that are firmly based on

empirical observation or a priori reasoning) and shows strong parallels with the "conventionalist" viewpoint, which did not become popular in science and philosophy until half a century later, when it numbered H. Hertz and POINCARÉ [VI.61] among its adherents.

Jacobi not only promoted new mathematical developments, but also studied the history of mathematics: he worked on ancient number theory, was the advisor for the historical parts of A. von Humboldt's great *Kosmos* (1845–62), and developed detailed plans for the publication of Euler's works.

Further Reading

Koenigsberger, L. 1904. *Carl Gustav Jacob Jacobi*. Festschrift zur Feier des hundertsten Wiederkehr seines Geburtstages. Leipzig: Teubner.

Helmut Pulte

VI.36 Peter Gustav Lejeune Dirichlet

b. Düren, French Empire (now Germany), 1805;

d. Göttingen, Germany, 1859

Number theory; analysis; mathematical physics; hydrodynamics; probability theory

The low level of mathematics education at German universities prompted Dirichlet to study in Paris, where he came into contact with the leading French mathematicians Lacroix, POISSON [VI.27], and FOURIER [VI.25], who particularly attracted him. In 1827 he took up a position at the University of Breslau. The following year he moved to Berlin, where he was appointed as a professor at the military academy and where he was also allowed to teach at the university. In 1831 he was made a professor at the university and from then on held positions at both institutions until 1855, when he was appointed as the successor to GAUSS [VI.26] at the University of Göttingen.

Dirichlet's primary interest was in number theory. His guiding star was Gauss's pioneering *Disquisitiones Arithmeticae* (1801)—the work that made number theory into a mathematical discipline—which he studied throughout his career. Dirichlet was not only the first mathematician to completely understand this work but he also became its interpreter, picking up its problems and improving its proofs, as well as developing its ideas.

With his very first publication, which appeared in 1825, Dirichlet came to international prominence. The paper, which dealt with Diophantine equations of the

form $x^5 + y^5 = Az^5$, yielded substantial results for the verification of FERMAT'S LAST THEOREM [V.12] for the case $n = 5$ (these results were used by LEGENDRE [VI.24] for a complete proof for that case some weeks later). In a paper published in 1837 Dirichlet came up with the new and revolutionary idea of applying analytical methods to number theory. He introduced expressions that are now known as *Dirichlet L-series*. These are infinite series of the form

$$L(s, \chi) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s}$$

where $\chi(n)$ is a *Dirichlet character* modulo k : that is, a complex-valued function on the integers that is totally multiplicative, in the sense that $\chi(ab) = \chi(a)\chi(b)$ for all a and b , and χ is periodic with period k and not identically zero. Using these L -series, Dirichlet showed that every arithmetic progression $\{an + b : n = 0, 1, \dots\}$, where a and b are relatively prime, contains infinitely many prime numbers. In two subsequent papers, published in 1838 and 1839, he used his new methods, among other things, to determine the formula of the class number of binary quadratic forms: that is, the number of proper classes of forms of given determinant. It is often said that these three papers mark the start of ANALYTIC NUMBER THEORY [IV.2].

Dirichlet also made important contributions to algebraic number theory, culminating in his UNIT THEOREM [III.65] for the Abelian group of units in an algebraic number field. These contributions, together with numerous others due to him (e.g., his *Schubfachprinzip*, or box principle; work on the law of biquadratic reciprocity; and results concerning Gaussian sums), were brought together in his influential *Vorlesungen über Zahlentheorie* (lectures on number theory), published by his former student DEDEKIND [VI.50] in 1863.

Inspired by his close contact with FOURIER [VI.25] during his student days in Paris, Dirichlet's other main interests were in analysis and mathematical physics, and in the connections between them. In a groundbreaking paper of 1829, Dirichlet not only gave the first strict proof of the convergence of a Fourier series under given conditions, but he also used new methods and concepts (e.g., his insight into the importance of conditional convergence of series; his Dirichlet function influencing the development of the concept of a function) that became classic and that served as a basis for countless nineteenth-century investigations on analysis. He also occupied himself with the determination of multiple integrals as well as with the expansion of

a function into spherical functions (*Kugelfunktionen*) and applied these results to problems in mathematical physics. His main contributions to mathematical physics include papers on the theory of heat, hydrodynamics, the gravitational attraction of an ellipsoid, the n -body problem, and potential theory. The first boundary-value problem (the "Dirichlet problem" of finding the solution of an elliptic partial differential equation in the interior of a given region that takes prescribed values on the boundary of the region) had already been handled by Fourier and others, but Dirichlet proved the uniqueness of the solution, while the DIRICHLET PRINCIPLE [IV.12 §3.5] (a method for solving boundary problems for ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS [IV.12 §2.5] by reducing them to VARIATIONAL PROBLEMS [III.96]) was introduced by him in lectures on potential theory, enhancing a method introduced by Gauss. Connected with Dirichlet's work on analysis was his contribution to probability and error theory, in particular his development of new methods for probabilistic limit theorems.

Dirichlet also influenced the further development of mathematics by his mathematical style, by the exactness and elegance of his proofs, and by his teaching. Together with his friend JACOBI [VI.35], he ushered in a new epoch of mathematical teaching at German universities by introducing lectures and seminars on the most recent research, and with him began the golden age of mathematics in Berlin. Although Dirichlet did not found his own mathematical school, his influence can be found in the work of Dedekind, Eisenstein, KRONECKER [VI.48], and RIEMANN [VI.49], among others.

Further Reading

- Butzer, P.-L., M. Jansen, and H. Zilles. 1984. Zum bevorstehenden 125. Todestag des Mathematikers Johann Peter Gustav Lejeune Dirichlet (1805–1859), Mitbegründer der mathematischen Physik im deutschsprachigen Raum. *Sudhoffs Archiv* 68:1–20.
- Kronecker, L., and L. Fuchs, eds. 1889–97. *G. Lejeune Dirichlet's Werke*, two volumes. Berlin: Reimer.

Ulf Hashagen

VI.37 William Rowan Hamilton

b. Dublin, 1805; d. Dublin, 1865

Calculus of variations; optics; dynamics; algebra; geometry

Hamilton was educated at Trinity College, Dublin. Shortly before graduating in 1827, he was appointed Professor of Astronomy and Royal Astronomer of Ireland, a post which he held for the remainder of his life.

His first paper, "Theory of systems of rays: part first" (1828), was written while he was still an undergraduate. In it he developed new methods for the study of foci and caustics produced by the reflection of light from curved surfaces. Hamilton developed his approach to optics over the following five years, publishing three substantial supplements to his original paper. He showed that the properties of an optical system are completely determined by a certain "characteristic function" that is a function of the initial and final coordinates of a ray of light and which measures the time of passage of light through the system. In 1832 he predicted that light falling at a certain angle on a biaxial crystal would be refracted to form a hollow cone of emergent rays. This prediction was verified by his friend and colleague Humphrey Lloyd.

Hamilton adapted his optical methods to the study of dynamics. In a paper "On a general method in dynamics" (1834), he showed that the dynamics of a system of attracting and repelling point particles is completely determined by a certain characteristic function, which satisfies a differential equation, today referred to as the HAMILTON-JACOBI EQUATION [IV.12 §2.1]. In a subsequent paper, "Second essay on a general method in dynamics" (1835), he introduced the *principal function* of a dynamical system, presented the equations of motion of such a system in HAMILTONIAN FORM [IV.16 §2.1.3], and adapted methods of perturbation theory to this setting.

Hamilton discovered the system of QUATERNIONS [III.78] in 1843. The fundamental equations of this system occurred to him in a flash of insight as he was walking along the bank of the Royal Canal, near Dublin, on October 16 of that year. Most of his subsequent mathematical work involved quaternions. It is not difficult to translate much of this work into the language of modern vector analysis, and indeed many of the basic concepts and results of vector algebra and analysis emerged from Hamilton's work on quaternions. Hamilton applied quaternion methods to the study of dynamics in a series of short papers published in the three years immediately following his discovery of quaternions. He also investigated a number of algebraic systems related to quaternions. However, most of his work with quaternions was concerned with their application

to the study of geometrical problems, and, in particular, to the study of surfaces of the second order and (especially in the final years of his life) to the study of the differential geometry of curves and surfaces. Much of this research is to be found in his two books *Lectures on Quaternions* (1853) and *Elements of Quaternions* (1866, published posthumously).

Further Reading

Hankins, T. L. 1980. *Sir William Rowan Hamilton*. Baltimore, MD: Johns Hopkins University Press.

David Wilkins

VI.38 Augustus De Morgan

b. Madura (now Madurai), India, 1806; d. London, 1871

Professor of Mathematics, University College London (1828–31, 1836–66); first president of the London Mathematical Society (1865–66)

De Morgan, a prolific author in many fields of mathematics and its history, made important and original contributions to the development of mathematical logic. He is particularly remembered for what we now call *de Morgan's laws*, which he first published in 1858 in a paper in the *Transactions of the Cambridge Philosophical Society*. The "laws" can be stated (using the notation of sets) as follows. If A and B are subsets of a set X , then $(A \cap B)^c = A^c \cup B^c$ and $(A \cup B)^c = A^c \cap B^c$, where " \cup " represents union, " \cap " represents intersection, and a superscript " c " denotes the complement with respect to X .

VI.39 Joseph Liouville

b. Saint Omer, France, 1809; d. Paris, 1882
Differentiation of arbitrary order; integration in closed form; Sturm-Liouville theory; potential theory; mechanics; differential geometry; doubly periodic functions; transcendental numbers; quadratic forms

Liouville was the leading French mathematician in the generation between CAUCHY [VI.29] and HERMITE [VI.47]. He taught analysis and mechanics at his alma mater, the École Polytechnique, until 1851, when he became professor at the Collège de France. Moreover, he was professor at the Sorbonne from 1857 and member of the Paris Academy of Sciences and the Bureau des Longitudes. In 1836 he founded the *Journal de Mathématiques Pures et Appliquées*, which exists to this day.

His wide-ranging research was often inspired by physics. For example, his early theory of differential operators of the form $(d/dx)^k$, where k is an arbitrary complex number, had its origin in Ampère's electrodynamics. Similarly, Sturm-Liouville theory, which he developed in around 1836 with his friend C. F. Sturm, was inspired by the theory of heat conduction. Sturm-Liouville theory deals with a linear self-adjoint second-order differential equation involving a parameter that must be chosen so that there exist nontrivial solutions (eigenfunctions) that satisfy given boundary-value conditions. Liouville's main contribution to this theory was a proof that an "arbitrary" function has a convergent "Fourier expansion" in terms of eigenfunctions. Sturm-Liouville theory was a major step toward a more qualitative theory of differential equations, and the first work on spectral theory of a general class of differential operators.

Liouville was the first to prove (in 1844) that there exist TRANSCENDENTAL NUMBERS [III.43], such as $\sum_{n=1}^{\infty} 10^{-n!}$. In a similar vein, in the 1830s he had already shown that there are elementary functions such as e^t/t whose integrals are not expressible in elementary (or closed) form, i.e., in terms of algebraic functions, exponentials, and logarithms. In particular he proved that the elliptic integrals are nonelementary.

Around 1844 Liouville suggested an entirely new approach to ELLIPTIC FUNCTIONS [V.34] (inverses of elliptic integrals), based on a systematic investigation of doubly periodic complex functions and in particular the observation that such a function must have singularities if it is not constant. When Cauchy heard of this theorem he immediately generalized it to the statement that any bounded complex analytic function must be a constant. Today this is called Liouville's theorem.

In mechanics, Liouville's name is connected with the theorem stating that the volume in phase space is constant when a mechanical system moves according to HAMILTON'S EQUATIONS [III.90 §2.1]. In fact, Liouville proved the constancy of a certain DETERMINANT [III.15] formed from the solutions of a general class of differential equations. It was JACOBI [VI.35] who pointed out that the theorem applied to Hamilton's equations, and Boltzmann who interpreted the determinant as the volume in phase space, and emphasized its importance in statistical mechanics.

Liouville made many other important contributions to mechanics and to potential theory. For example, Jacobi had postulated that when the angular momentum of a fluid planet revolving around an axis is high

enough, there are two shapes that are in equilibrium in their rotating frames of reference: an ellipsoid of revolution and an ellipsoid with three different axes. Liouville showed that Jacobi was right, and moreover proved the surprising result that only the latter figure is in stable equilibrium. Liouville published only the result, leaving the verification to Lyapunov and POINCARÉ [VI.61] (at least if the angular momentum is not too large).

The first mathematician to recognize the significance of GALOIS'S [VI.41] theory of SOLVABILITY OF EQUATIONS [V.23], Liouville did a great service to algebra when he published some of Galois's most important papers in his journal.

Further Reading

Lützen, J. 1990. *Joseph Liouville 1809–1882: Master of Pure and Applied Mathematics*. Studies in the History of Mathematics and Physical Sciences, volume 15. New York: Springer.

Jesper Lützen

VI.40 Eduard Kummer

b. Sorau (now Zary, Poland), 1810; d. Berlin, 1893

Gymnasium teacher, Liegnitz (now Legnica, Poland) 1832–42; Professor of Mathematics: Breslau (now Wrocław, Poland) 1842–55, Berlin 1855–82

Kummer's early research was in function theory, in which he made an important contribution to the theory of the (generalized) hypergeometric series (a power series in which the ratios of successive coefficients are rational functions). Surpassing earlier work of GAUSS [VI.26], Kummer not only provided a systematic account of solutions to the hypergeometric differential equation

$$x(x-1) \frac{d^2 y}{dx^2} + (c - (a+b+1)x) \frac{dy}{dx} - aby = 0,$$

where a , b , and c are constants, but also made the connection between the hypergeometric functions and newer functions in analysis, such as the ELLIPTIC FUNCTIONS [V.34].

After moving to Breslau, Kummer started doing research in number theory, the field in which he achieved his greatest success: the creation of the theory of "ideal prime factors" (1845–47). Although Kummer's theory is often described as an early contribution to the theory of IDEALS [III.83 §2], his algorithmic approach was very

different from that later followed by DEDEKIND [VI.50]. Kummer's original goal had been to generalize the LAW OF QUADRATIC RECIPROCITY [V.31] to higher powers, and he succeeded in this in 1859. An additional consequence of this research was that he managed to prove FERMAT'S LAST THEOREM [V.12] for all prime exponents (and hence, since it was known for fourth powers, all exponents) less than 100.

In the third phase of his career Kummer turned to algebraic geometry. Continuing the work of HAMILTON [VI.37] and JACOBI [VI.35] on ray systems and geometric optics, he was led to the discovery of the quartic surface with sixteen nodal points, which is now named after him.

VI.41 Évariste Galois

b. Bourg-la-Reine, France, 1811; d. Paris, 1832

Theory of equations; theory of groups; Galois theory; finite fields

Galois studied at home until he was eleven years old, then entered the Collège Louis-le-Grand in Paris, where he stayed for six years. He had, and gave his teachers, a difficult time there, but excelled in mathematics, in which he read advanced work of LAGRANGE [VI.22], GAUSS [VI.26], and CAUCHY [VI.29] alongside standard texts of the time. He attempted the entrance examination for the École Polytechnique prematurely in June 1828, but failed. In July 1829, after his father's suicide, Galois was again rejected by the École Polytechnique. He entered the École Préparatoire (later known as the École Normale Supérieure) in October 1829 but was expelled in December 1830 for unacceptable behaviour arising from political disagreements with the authorities. Arrested on Bastille Day (14 July) 1831, he spent the next eight months in prison for flouting authority again. He emerged at the end of April 1832 but somehow got himself challenged to a duel. On 29 May he edited his manuscripts and wrote a summary of his discoveries in a letter to his friend Auguste Chevalier. The duel took place the next morning and he died on 31 May 1832. Much has been written about him. But a man who dies so young leaves little real evidence for historians to work with, however rich his story, and most of his biographers have allowed romantic invention to colour their accounts of his life.

There are four main papers in Galois's mathematical works (and a number of smaller and less important items). The first to be published was "Sur la théorie des nombres," which appeared in April 1830 and contains the theory of Galois fields. These are analogues of the

complex numbers obtained by adjoining to the integers modulo a prime number p a root of an irreducible polynomial congruence modulo p . The paper contains most of the basic features of what later became the theory of finite fields.

In the letter to Chevalier written on the eve of the duel, Galois mentions three memoirs. The first, now known as the *Premier Mémoire*, is a manuscript entitled "Sur les conditions de résolubilité des équations par radicaux." Galois submitted work on the theory of equations to the Paris Academy on 25 May and 1 June 1829, but this is now lost and it seems quite possible that Galois withdrew it on the advice of Cauchy (to whom it had been given to referee) in January 1830. In February 1830 he resubmitted his work in competition for the Grand Prix de Mathématiques, but his manuscript was unfortunately and mysteriously lost on the death of FOURIER [VI.25] (and the prize was awarded jointly to ABEL [VI.33], posthumously, and JACOBI [VI.35]). Encouraged to do so by POISSON [VI.27] he submitted his ideas to the Academy for a third time in January 1831. It is this third submission (which was read by the Academy referees, Poisson and Lacroix, and rejected on 4 July 1831) that survives as the manuscript of the *Premier Mémoire*. This is the remarkable work in which he introduced what is now called the Galois group of an equation and showed how solubility of the equation in terms of radicals could be precisely characterized by a property of the group. It was the *Premier Mémoire* which turned the theory of equations into what is now called GALOIS THEORY [V.24].

The *Second Mémoire* also exists. Galois never completed it, however, nor is it all correct. Nevertheless, it is an exciting document that focuses on aspects of what is now recognized as the theory of groups. Its main theorem is (in group-theoretic language) that every primitive soluble permutation group has degree a power of a prime number and may be represented as a group of affine transformations over the prime field \mathbb{F}_p . It also contains an incomplete study of two-dimensional linear groups over \mathbb{F}_p . The *Troisième Mémoire*, which he described as being on the theory of integrals and ELLIPTIC FUNCTIONS [V.34], has never been found.

Galois's main work—comprising the paper "Sur la théorie des nombres," the *Premier Mémoire*, the *Second Mémoire*, and the letter to Chevalier—was finally published by LIOUVILLE [VI.39] in 1846. A critical edition by Bourgne and Azra, including every known fragment of Galois's writing, was published in 1962.

PUP: this is the article where UK spelling was approved for the (ultra-pedantic) author, so I have also kept UK date style as it was his strong preference. OK?

Galois's legacy is enormous. His ideas led directly to "abstract algebra" (see [II.3 §6]): when the abstract notion of field developed later in the nineteenth century, it turned out that most of the theory of finite fields had already been anticipated in that first paper; Galois theory developed directly out of the material in the *Premier Mémoire*; and the theory of groups developed from the ideas in the *Premier Mémoire* and the *Second Mémoire* together with a series of papers published by Cauchy in 1845.

Further Reading

- Bourgne, R., and J.-P. Azra, eds. 1962. *Écrits et Mémoires Mathématiques d'Évariste Galois*. Paris: Gauthiers-Villars.
 Edwards, H. M. 1984. *Galois Theory*. New York: Springer.
 Taton, R. 1983. Évariste Galois and his contemporaries. *Bulletin of the London Mathematical Society* 15:107–18.
 Toti Rigatelli, L. 1996. *Évariste Galois 1811–1832*, translated from the Italian by J. Denton. Basel: Birkhäuser.

Peter M. Neumann

VI.42 James Joseph Sylvester

b. London, 1814; d. London, 1897
Algebra

As a Jew, Sylvester could neither take the degree he earned at St John's College, Cambridge, in 1837 nor compete for positions at England's Anglican universities. This effectively forced him down a convoluted path toward his personal goal of a career as a research mathematician. He worked as an actuary in London in the 1840s and 1850s before qualifying as a lawyer by passing the English Bar. He was unemployed for some six years in the 1870s, but held professorships at various times, both of natural philosophy and of mathematics, in England and in the United States. Most notably, Sylvester served as the first Professor of Mathematics at Johns Hopkins University in Baltimore, Maryland, from 1876 to 1883 and, thanks to an 1871 law that finally made it possible for non-Anglicans to hold professorships at Oxbridge, was eligible for and won the appointment as Oxford's Savilian Professor of Geometry in 1883. He held the Oxford chair until ill health forced his retirement in 1894. The program Sylvester set up at Johns Hopkins established his pivotal place in the history of American research-level mathematics, while his mathematical accomplishments had garnered him an international reputation as early as the 1860s.

Sylvester entered the research arena in the late 1830s with work on the problem of determining when two polynomial equations have a common root. This naturally led not only to questions in the theory of determinants but also to an explicit, pioneering, and self-consciously algebraic analysis of the intermediate expressions that arise in Charles François Sturm's algorithm for determining the number of real roots of a polynomial equation that lie between two given real numbers (1839, 1840). Sylvester followed this up with what he called the dialytic method of elimination: a new criterion in terms of DETERMINANTS [III.15] for detecting whether two polynomial equations have a common root (1841).

His next major research push came in the 1850s when, together with CAYLEY [VI.46], he formulated a theory of invariants. This involved an associated and slightly more general theory of "covariants." More concretely, given a binary form of a particular degree, Sylvester and Cayley devised techniques both for explicitly finding invariants and covariants of that form and for determining algebraic relations, or "syzygies," between them. Sylvester tackled these questions in two important papers: "On the principles of the calculus of forms" (1852) and "On a theory of the syzygetic relations of two rational integral functions" (1853). In the latter, he proved, among other results, Sylvester's law of inertia: if $Q(x_1, \dots, x_n)$ is a real QUADRATIC FORM [III.75] of rank r , then there exists a (real) nonsingular linear transformation that takes Q to $x_1^2 + \dots + x_p^2 - x_{p+1}^2 - \dots - x_r^2$, where p is uniquely determined.

Sylvester surprised the mathematical world in 1864 and 1865 with the first proof of NEWTON's [VI.14] rule (Newton had only stated it) for determining bounds on the number of positive and negative roots of a polynomial equation. However, he then entered a fallow period that ended only with his move to Baltimore. While there, he returned to invariant theory, and specifically to the problem of inductively determining, for binary forms first of degree 2 then of degree 3 then of degree 4, etc., the number of covariants in a minimum generating set associated with the form. In 1868, Paul Gordan had proved that this number is always finite and, in so doing, had proved wrong an earlier result of Cayley, who claimed to have shown that a minimum generating set of covariants for the binary quintic form (that is, the binary form of degree 5) was infinite. By 1879, Sylvester had explicitly calculated minimum generating sets of covariants associated with binary forms of degrees two through ten. He had also succeeded in recognizing and

filling (1878) a critical gap in the proof that Cayley had given of a theorem on the maximal number of linearly independent covariants associated with a binary form of any given degree.

Sylvester was the founding editor of the *American Journal of Mathematics*, and indeed much of this invariant-theoretic work, as well as results on partitions (1882), on rational points on a cubic curve (1879–80), and on matrix algebras (1884), appeared there.

Further Reading

Parshall, K. H. 1998. *James Joseph Sylvester: Life and Work in Letters*. Oxford: Clarendon.

———. 2006. *James Joseph Sylvester: Jewish Mathematician in a Victorian World*. Baltimore, MD: Johns Hopkins University Press.

Sylvester, J. J. 1904–12. *The Collected Mathematical Papers of James Joseph Sylvester*, four volumes. Cambridge: Cambridge University Press. (Also reprint edition published in 1973. New York: Chelsea.)

Karen Hunger Parshall

VI.43 George Boole

b. Lincoln, England, 1815; d. Cork, Republic of Ireland, 1864
Boolean algebra; logic; operator theory; differential equations; difference equations

Boole, who never attended secondary school, college, or university, was almost entirely self-taught. His father was a poor shoemaker who was more interested in building telescopes and scientific instruments than making shoes—the result being that his business failed and Boole had to leave school at the age of fourteen and take a job as a junior teacher to support his parents, sister, and two brothers. By the age of ten he had mastered Latin and Ancient Greek, and by the age of sixteen he could read and speak French, Italian, Spanish, and German fluently. From his father he got a love of mechanics, physics, geometry, and astronomy, and together they built functioning scientific instruments. Boole then turned to mathematics and by the age of twenty he was publishing original research in calculus and linear systems. He wrote two seminal papers on linear transformations (1841, 1843), which provided the starting point for invariant theory, but he left it to others such as CAYLEY [VI.46] and SYLVESTER [VI.42] to develop the subject. In 1844 he was awarded the Royal Society's Gold Medal for his paper on operators in analysis, the first gold medal for mathematics to be

presented by the society. The paper was important not only because it contained (arguably for the first time) a clear definition of the concept of an OPERATOR [III.52], but also because of the influence it had on Boole's subsequent ideas. An operator, for Boole, was an operation of the calculus, such as differentiation (which he denoted by D), considered as an object in its own right. There was an explicit similarity between the laws he derived for functions of D and the laws of his algebra of logic, which we shall discuss below.

At one time Boole had hoped to become a clergyman but family circumstances prevented this. His reverence for creation made him interested in the workings of the human mind, which he regarded as God's greatest accomplishment. He longed, as Aristotle and LEIBNIZ [VI.15] had before him, to explain how the brain processes information and to express this information in mathematical form. In 1847 he published a book entitled *A Mathematical Analysis of Logic* in which he took the first steps toward achieving his goal, but the book did not have a wide circulation and so made very little impact on the mathematical world.

In 1849 Boole was appointed Professor of Mathematics at Queen's College, Cork. It was there that he rewrote and expanded his ideas in a book entitled *An Investigation of the Laws of Thought* (1854), in which he introduced a new type of algebra, an algebra of logic, which evolved into what we now call Boolean algebra. From his earlier study of languages, he realized that there were mathematical structures concealed in everyday speech. For example, the class of European men, together with (i.e., union) the class of European women, is the same as the class of European men and women. By using letters to represent a class, or set, of objects, he could write the above as $z(x + y) = zx + zy$, where the letters x , y , and z represent the class of men, the class of women, and the class of all Europeans, respectively. Here addition is to be understood as union, at least for disjoint classes like men and women, and multiplication is to be understood as intersection.

The principal laws of Boole's algebra are commutativity, distributivity, and the law which he called the "fundamental law of duality" and which is represented by $x^2 = x$. This law can be interpreted by observing that the class of all white sheep intersected with the class of all white sheep is still the class of all white sheep. Unlike his other laws, all of which apply to ordinary numerical algebra, this law applies to numerical algebra only when x is 0 or 1.

Boole broke with traditional mathematics by showing that the study of well-defined classes or sets of objects is capable of precise mathematical interpretation and is indeed fundamental to mathematical analysis. In simple cases, his approach also reduces classical logic to symbolic mathematical form. Using the symbols 0 and 1 to denote “nothing” and “universe” respectively, and denoting the complement of the class x by $1 - x$, he derived (from the law of duality) the law $x(1 - x) = 0$, which represents the impossibility of an object simultaneously possessing and not possessing a given property, otherwise known as the principle of contradiction. Boole also applied his calculus to the theory of probability.

Boole’s algebra lay dormant until 1939, when Shannon discovered that it was the appropriate language for describing digital switching circuits. Boole’s work thus became an essential tool in the modern development of electronics and digital computer technology.

Boole also made several other contributions to mathematics: differential equations, difference equations, operator theory, calculus of integrals, etc. His textbooks on differential equations (1859) and finite differences (1860) include much of his original research and are still in print today, but he is best remembered as the father of symbolic logic and one of the founders of computer science.

Further Reading

MacHale, D. 1983. *George Boole, His Life and Work*. Dublin: Boole Press.

Des MacHale

VI.44 Karl Weierstrass

b. Ostenfelde, Germany, 1815; d. Berlin, 1897
Analysis

Weierstrass began his career studying finance and administration at the University of Bonn but his real interest was mathematics and he did not complete his course. He qualified as a teacher and taught in gymnasias for fourteen years. The turning point in his life occurred when, at the age of almost forty, he published a ground-breaking paper on Abelian functions, in which he solved the problem of inversion of hyperelliptic integrals. Shortly afterward he was offered a position at the University of Berlin. He demanded of himself the very strictest standards, with the result that he published

little. His ideas, and his reputation, spread through his excellent lectures, which drew students and established mathematicians from around the world.

Weierstrass has been described as the “father of modern analysis.” He contributed to all branches of the subject: calculus, differential and integral equations, CALCULUS OF VARIATIONS [III.96], infinite series, elliptic and Abelian functions, and real and complex analysis. His work is characterized by attention to foundations and by scrupulous logical reasoning. “Weierstrassian rigor” has come to denote rigor of the strictest standard.

Calculus in the seventeenth and eighteenth centuries was heuristic, lacking logical foundations. The nineteenth century ushered in a rigorous spirit in mathematics which included an examination of the foundation of various fields of mathematics. CAUCHY [VI.29] initiated this process in calculus in the 1820s. But there were several major foundational problems with his approach: verbal definitions of limit and continuity; frequent use of infinitesimals; and intuitive appeal to geometry in proving the existence of various limits.

Weierstrass and DEDEKIND [VI.50] (among others) determined to remedy this unsatisfactory situation, with the goal of establishing theorems in a “purely arithmetic” manner, as Dedekind put it. To that end, Weierstrass gave precise ϵ - δ definitions of LIMIT [I.3 §5.1] and CONTINUITY [I.3 §5.2] (those we still use today), thus banishing infinitesimals from analysis (until ROBINSON [VI.95] some hundred years later). He also defined the real numbers based on the rationals (although Dedekind’s and CANTOR’s [VI.54] approaches proved more accessible). He was thereby largely responsible for the “arithmetization of analysis” (a term coined by KLEIN [VI.57]). Among his remarkable contributions to real analysis are his introduction of uniform convergence (introduced independently by P. L. Seidel) and his example of an everywhere-continuous and nowhere-differentiable function (Cauchy and his contemporaries believed that a continuous function was differentiable except possibly at isolated points).

Both RIEMANN [VI.49] and Weierstrass (succeeding Cauchy) founded complex function theory, but they had fundamentally different approaches to the subject. Riemann’s global, geometric conception was based on the notion of a RIEMANN SURFACE [III.81] and on the DIRICHLET PRINCIPLE [IV.12 §3.5], while Weierstrass’s local algebraic theory was grounded in power series and ANALYTIC CONTINUATION [I.3 §5.6]. “The more I

ponder the principles of function theory—and I do so incessantly—the more I am convinced that it must be founded on simple algebraic truths....” he asserted in a letter to H. A. Schwartz. He severely criticized the Dirichlet principle for being mathematically not well-grounded, and produced a counterexample, after which his approach to complex analysis became dominant until the early twentieth century. Klein commented on Weierstrass’s general approach to mathematics: “[He] is first of all a logician; he proceeds slowly, systematically, step-by-step. When he works, he strives for the definitive form.”

Weierstrass’s name is attached to various concepts and results, among them the *Weierstrass approximation theorem*, which says that a continuous function can be uniformly approximated by polynomials; the *Bolzano–Weierstrass theorem*, which states that every infinite, bounded set of real numbers has a limit point; the *Weierstrass factorization theorem*, which gives the representation of an entire function in terms of an infinite product of “prime functions”; the *Casorati–Weierstrass theorem*, which says that in every neighborhood of an isolated essential singularity an analytic function takes values arbitrarily close to any assigned complex number; the *Weierstrass M-test*, which deals with the comparison of series for convergence; and the *Weierstrass \wp -function*, an example of an ELLIPTIC FUNCTION [V.34] of order 2.

Weierstrass was most proud of his work on Abelian functions, and much of his fame in the nineteenth century rested on it. His results in this field are, however, less significant today. For us, his main legacy is his unrelenting insistence on maintaining high standards of rigor and seeking the fundamental ideas underlying mathematical concepts and theories.

Further Reading

Bottazzini, U. 1986. *The Higher Calculus: A History of Real and Complex Analysis from Euler to Weierstrass*. New York: Springer.

Israel Kleiner

VI.45 Pafnuty Chebyshev

b. Okatovo, Russia, 1821; d. Saint Petersburg, Russia, 1894

Assistant, Extraordinary then full Professor of Mathematics, Saint Petersburg (1847–82); Artillery Committee (1856); Scientific Committee of the Ministry of Education (1856)

Fascinated by Watt’s parallelogram (the linkage used in steam engines) and the problem of converting circular motion into rectilinear motion, Chebyshev embarked on a deep study of the theory of hinge mechanisms. In particular, he sought the linkage that would produce the minimum deviation from a straight line over a given range. This corresponds to the mathematical problem of finding, from among the class of functions chosen to approximate a given function, the one with the smallest absolute error for all specified values of the argument. It was in this context, in particular considering the approximation of functions by polynomials, that Chebyshev discovered the polynomials now named after him (see [III.87]). These polynomials were first published in his memoir “Théorie des mécanismes connus sous le nom de parallélogrammes” (1854), and they marked the beginning of his important contributions to the theory of orthogonal polynomials.

Chebyshev polynomials of the first kind are defined by $T_n(\cos \theta) = \cos(n\theta)$, for $n = 0, 1, 2, \dots$. These polynomials also satisfy the recurrence relation $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$, where $T_0(x) = 1$ and $T_1(x) = x$. Chebyshev polynomials of the second kind satisfy $U_n(\cos \theta) = \sin((n+1)\theta)/\sin \theta$ and the recurrence relation $U_{n+1}(x) = 2xU_n(x) - U_{n-1}(x)$, where $U_0(x) = 1$ and $U_1(x) = 2x$.

Chebyshev also had a significant impact on number theory, coming close to proving the PRIME NUMBER THEOREM [V.29]. In probability he is remembered for Chebyshev’s inequality, a result that is simple but has innumerable applications.

VI.46 Arthur Cayley

b. Richmond, England, 1821; d. Cambridge, England, 1895
Algebra; geometry; mathematical astronomy

At the beginning of his career in the 1840s, Cayley laid down subjects that informed much of his later research. The novelties of his very first undergraduate paper, “On a theorem in the geometry of position” (1841), are the now-standard notation for DETERMINANTS [III.15] of arrays set between vertical lines and the introduction of the *Cayley–Menger determinant*. Following HAMILTON’s [VI.37] discovery of the QUATERNIONS [III.78] (1843), Cayley expressed rotations in three-dimensional space via the succinctly expressed mapping $x \rightarrow q^{-1}xq$, a result that led him to the *Cayley–Klein parameters*. He outlined the nonassociative system of the octaves (CAYLEY NUMBERS [III.78]), the intersection of curves (the *Cayley–Bacharach*

theorem), and a dual curve called the *Cayleyan*. In major papers, he described a theory of multilinear determinants and ELLIPTIC FUNCTIONS [V.34] as doubly infinite products. In concert with George Salmon he investigated the famous twenty-seven lines that lie in a cubic surface. The most important studies among his juvenilia, though, were his first steps in invariant theory (1845, 1846), the field in which his reputation was made.

Between 1849 and 1863, years spent as a qualified London barrister, Cayley broadened his range, but unlike other gentlemen of science who roamed across a multitude of subjects, he restricted his activity exclusively to mathematics. This was mostly pure mathematics. He generalized PERMUTATION GROUPS [III.70] using the calculus of operations as a basis, and he saw that not only were matrices useful as a notational device, but they also constituted a study in their own right. Not generally an excitable person, at the point of discovery he declared the *Cayley-Hamilton theorem* as “very remarkable” and generations of mathematicians have shared his delight. Matrix algebra was used in his solution of the *Cayley-Hermite problem*, which required a description of those linear transformations that leave a bilinear form invariant. A special case of the solution gives rise to the *Cayley orthogonal transform* $(I - T)(I + T)^{-1}$. The links between quaternions, matrices, and group theory that he observed in the 1850s are indicative of his concern for the organization of mathematics.

In the 1850s, Cayley set in motion his famous memoirs on *quantics*, a term he coined for algebraic forms, now referred to as multilinear homogeneous algebraic forms. He discovered *Cayley’s formula* for the general form of covariants of binary forms and *Cayley’s law* for counting them. In the *Sixth Memoir* (1859), he demonstrated that EUCLIDEAN GEOMETRY [I.3 §6.2] was part of PROJECTIVE GEOMETRY [I.3 §6.7] rather than the converse. The idea of a projective metric (*Cayley’s absolute*) was seen by KLEIN [VI.57] in the 1870s as the unifying conceptual idea for classifying non-Euclidean geometries.

For twenty-five years, from 1858, he was the editor of the *Monthly Notices of the Royal Astronomical Society*. In astronomy he contributed to the theory of elliptic planetary motion, calculatory work that demanded an assiduous attention to detail. His work on the lunar theory was noteworthy, and in one long calculation he helped to settle an Anglo-French controversy by verifying the correct value for the secular acceleration of

the Moon, which had been established by John Couch Adams in 1853.

Cayley returned to the academic world in 1863 as the founding Sadleirian Professor of Pure Mathematics at Cambridge. In 1868 Paul Gordan startled invariant theorists by proving that invariants and covariants of a binary quantic could be expressed in terms of a *finite* basis. This contradicted an earlier result of Cayley’s but, undaunted, he completed his series with a listing of the irreducible invariants and covariants of the binary form of order 5 (the binary quintic), and their connecting syzygies.

Many developments in pure mathematics can be traced back to his minor notes of the 1870s and 1880s, including the theory of knots, fractals, dynamic programming, and group theory (the well-known *Cayley’s theorem*). In graph theory, the number of distinct labeled trees with n nodes being n^{n-2} is known as *Cayley’s graph theorem*. He brought his theoretical knowledge of graphical trees to bear on the problem of counting isomers in organic chemistry, thus prompting questions about the actual existence of certain chemical compounds that have since been discovered in many instances by chemists. In the last decade of his life, Cayley set about the task that gave him an important line of contact with today’s mathematicians: the publication of his *Collected Mathematical Papers* in thirteen large volumes by Cambridge University Press.

Further Reading

Crilly, T. 2006. *Arthur Cayley: Mathematician Laureate of the Victorian Age*. Baltimore, MD: Johns Hopkins University Press.

Tony Crilly

VI.47 Charles Hermite

b. Dieuze, Moselle, France, 1822; d. Paris, 1901

Analysis (elliptic functions, differential equations);

algebra (invariant theory, quadratic forms); approximation theory

Like many who aspired to enter the École Polytechnique, Hermite undertook special preparatory classes, in his case at Lycée Henri IV and Lycée Louis-le-Grand. He began to study serious mathematics, immersing himself in the work of LAGRANGE [VI.22] and LEGENDRE [VI.24], and became interested in the solution of equations by radicals. Admitted to the École Polytechnique in 1842, by the end of that year he had completed his first significant original work. This extended results

of JACOBI [VI.35] in the theory of ELLIPTIC FUNCTIONS [V.34]. He sent these to Jacobi, who responded very positively. This achievement both brought him recognition in Paris and initiated a correspondence with Jacobi on elliptic functions and number theory that launched Hermite's career.

Hermite nonetheless struggled to find a position commensurate with his abilities, and for almost a decade survived on teaching assistant and examiner jobs around Paris. Hermite's work turned to number theory, in particular the arithmetic of quadratic forms, where he followed GAUSS [VI.26] and Lagrange in studying when one form can be reduced to another by a linear transformation. It was in this context that the HERMITIAN MATRICES [III.52 §3] named after him arose. Hermite was interested in invariants of quadratic forms, and also applied his work to the problem of location of roots of polynomials. As a result of these efforts, in 1856 he was appointed to the Paris Academy of Sciences, with LIOUVILLE [VI.39] and CAUCHY [VI.29] supporting him. This appointment was quickly followed by Hermite's 1858 discovery of a means to express the solutions of the general fifth-degree polynomial equation in terms of elliptic functions, which earned him widespread international recognition.

Finally obtaining a professorship at the Faculty of Science in Paris in 1869, Hermite became an influential mentor for a generation of mathematicians, his best-known protégés including J. Tannery, POINCARÉ [VI.61], E. Picard, P. Appell, and E. Goursat. Hermite's dynastic connections are also impressive: his brother-in-law, Joseph Bertrand, was permanent secretary of the Paris Academy of Sciences, Picard was his son-in-law, Appell married Bertrand's daughter, and their daughter married BOREL [VI.70]. His advocacy of improved international communication led to German work becoming much better known in France than it had been previously. During this period, he obtained a proof of the TRANSCENDENCE [III.43] of "e" using CONTINUED-FRACTION [III.22] methods based on earlier research in approximation theory (which had included the invention of the Hermite polynomials). His influence in the mathematical community was strong until his death.

Further Reading

Picard, É. 1901. L'œuvre scientifique de Charles Hermite. *Annales Scientifiques de l'École Normale Supérieure* (3) 18: 9–34.

VI.48 Leopold Kronecker

b. Liegnitz, Silesia, today Poland, 1823; d. Berlin, 1891
Algebra; number theory

One of the dominant mathematicians of the second half of the nineteenth century, Kronecker is best known today for his constructivist views and his contributions to number theory. After finishing his Ph.D. under the supervision of DIRICHLET [VI.36] in 1845, Kronecker left Berlin and mathematics in order to manage a family estate and to wind up his father-in-law's banking business. These activities left him wealthy and free to return to Berlin and concentrate on mathematics without holding an academic position. In 1855, Kronecker's former school teacher and closest scientific friend, EDUARD KUMMER [VI.40], also came to Berlin and stayed there until his death in 1893. In 1861, Kronecker became a member of the Berlin Academy of Sciences and started teaching courses at Berlin University. Kronecker valued the exchange with his Berlin colleagues (especially Kummer and WEIERSTRASS [VI.44]) highly, until a quarrel arose between Kronecker and Weierstrass in the 1870s, which drove Weierstrass to bitter, even anti-Semitic, complaints to others about Kronecker. After Kummer's retirement in 1883, Kronecker occupied Kummer's chair and stepped up his teaching activities as well as the frequency of his publications. This last active period was cut short when he died, shortly after the death of his wife.

Renowned for the originality of his mathematical insight, Kronecker became increasingly influential through the 1860s and 1870s. In 1868, he was offered the chair at Göttingen formerly held by GAUSS [VI.26], and was elected to the Paris Academy. After the Franco-Prussian war of 1870–71, he was invited to recommend mathematicians for the newly opened German university in Strasbourg; and in 1880 he became the managing editor of the *Journal für die reine und angewandte Mathematik* (otherwise known as *Crelle's Journal*). He was often criticized for incomplete, unpublished, or incomprehensible proofs—JORDAN [VI.52] spoke of his colleagues' "envy and despair" with regard to his results. Only in his later years was he explicit about his constructivist methodology. This constituted at least part of the quarrel with Weierstrass, and later prompted HILBERT [VI.63] to call Kronecker a "Verbotsdiktator" ("forbidding dictator"). Generally affable and hospitable, Kronecker was tough in defending his mathematical ideas and his claims to priority.

Tom Archibald

In his first works on solvable algebraic equations (in the early 1850s), he claimed not only the so-called *Kronecker–Weber theorem* (in today’s formulation: every finite Galois extension of the rational numbers with Abelian GALOIS GROUP [V.24] lies in a field generated by roots of unity; the first correct proof of it was given by Hilbert in 1896), but also a generalization to Abelian extensions of imaginary quadratic fields, which he later called his “liebster Jugendtraum” (“dearest dream of his youth”). This dream, which was incorrectly translated by Hilbert into his twelfth problem in 1900, is today part of CLASS FIELD THEORY [V.31] and the theory of complex multiplication. Such connections between algebra, analysis, and arithmetic continued to pervade Kronecker’s later work. Important results of Kronecker include class number relations and limit formulas in the theory of ELLIPTIC FUNCTIONS [V.34], the structure theorem for finitely generated Abelian groups, and a theory of bilinear forms.

In the late 1850s, Kronecker began to work on algebraic number theory, but only in 1881 did he publish his “Grundzüge einer arithmetischen Theorie der algebraischen Grössen,” dedicated to Kummer on the fiftieth anniversary of his doctorate. This mathematical testament contains an (incomplete) exposition of a unified arithmetical theory of algebraic numbers and algebraic functions. As a research program, it adumbrates important aspects of class field theory as well as of an arithmetico-geometric theory in dimensions higher than one. Kronecker’s concept of “divisor” is equivalent to Dedekind’s notion of “ideal” in the case of Dedekind domains, but is more restricted in the general case. Several mathematicians, such as H. Weber, K. Hensel, and G. König, took up the “Grundzüge” in their own work.

On a more general level, Kronecker asked for the complete arithmetization of pure mathematics, i.e., for the effective finitary reduction of pure mathematics to the notion of positive integer. For this, he propagated the introduction of indeterminates and equivalence relations, a method which he traced back to Gauss. In the case of a finite extension of the rational numbers, for instance, Kronecker is explicitly working with polynomials modulo an irreducible equation $f(x) = 0$, rather than adjoining a root of it.

Further Reading

Kronecker, L. 1895–1930. *Werke*, five volumes. Leipzig: Teubner.
 Vlăduț, S. G. 1991. *Kronecker’s Jugendtraum and Modular Functions*. New York: Gordon & Breach.

Norbert Schappacher and Birgit Petri

VI.49 Georg Friedrich Bernhard Riemann



b. Breselenz, near Dannenberg, Germany, 1826;
 d. Selasca, Italy, 1866
*Real and complex analysis; differential equations;
 differential geometry; heat distribution; number theory;
 propagation of shock waves; topology*

Riemann was born into a poor pastor’s family and studied mathematics at Göttingen, eventually becoming a professor there. His health broke in 1862 and he died near Lake Maggiore, Italy, of pleurisy at the age of thirty-nine.

No mathematician is more associated with the mid-nineteenth-century transition from algorithmic to conceptual thought than Riemann. His doctoral thesis of 1851, and still more his paper on Abelian functions (1857), promoted the view that a HOLOMORPHIC FUNCTION [I.3 §5.6] is properly defined by the CAUCHY–RIEMANN EQUATIONS [I.3 §5.6] and is to be studied through a close connection with the theory of HARMONIC FUNCTIONS [IV.24 §5.1]. In his thesis he sketched a proof of the remarkable RIEMANN MAPPING THEOREM [V.37]. This states that if X and Y are any two simply connected open subsets in the complex plane, neither of which is the whole plane, then there is a holomorphic map from one to the other with a holomorphic inverse. For example, if you draw any closed curve in the plane that does not intersect itself, and let D be

the region inside the curve, then D is biholomorphically equivalent to the open unit disk. In the 1857 paper he defined RIEMANN SURFACES [III.81], showed how to analyze them topologically, and outlined the Riemann inequality which his student Gustav Roch improved to the RIEMANN-ROCH THEOREM [V.34] in 1864. (The Riemann-Roch theorem, which is of great importance in algebraic geometry as well as complex analysis, determines the dimension of the space of meromorphic functions on a given Riemann surface with a prescribed number of poles.) In 1857 he extended the theory of differential equations, specifically the important case of the hypergeometric equation, to complex functions. In 1859 he used deep, new ideas from complex function theory to study the (Riemann) zeta function and proposed his celebrated conjecture, THE RIEMANN HYPOTHESIS [IV.2 §3], concerning the location of the complex zeros of this function. The conjecture remains unsolved to this day.

These ideas enabled mathematicians to study complex functions on domains other than the plane and subsets of the plane. They opened the way to a geometric study of algebraic functions and algebraic curves, and proved to be decisive in the study of the integrals of algebraic functions (the theories of Abelian functions and theta functions of several variables). Investigations of the Riemann zeta function led not only to the discovery of new properties of classes of complex functions, but more recently to the use of zeta functions of many other kinds in other branches of mathematics, including dynamics.

In 1854 Riemann, inspired by his mentor DIRICHLET [VI.36], formulated the concept of the RIEMANN INTEGRAL [I.3 §5.5], which permitted him to do profound work on the convergence of trigonometric series. Dirichlet had been able to prove that a real function was correctly represented by a Fourier series, but only under very restrictive conditions. This left open the questions of what sorts of functions did not satisfy these conditions and how could they be studied. Riemann reformulated the concept of the integral and was able to show that it is not just the continuity of a function and the ways in which it may fail to be continuous that affect the accuracy of its Fourier series representation, but the nature of its oscillations. The Riemann integral remained the dominant definition of the integral until it was replaced by the LEBESGUE INTEGRAL [III.57] after 1902, which is better adapted to capturing the way the behavior of a function affects its Fourier series.

In a lecture, also given in 1854 (but published posthumously in 1868), he entirely reformulated geometry as being about spaces (sets of points, which he called MANIFOLDS [I.3 §6.9]) with a RIEMANNIAN METRIC [I.3 §6.10] (an appropriate concept of distance) and argued that the geometric properties of a space were its intrinsic ones. He noted that there are three constant-curvature spaces in two dimensions and showed how the idea of constant curvature can be extended to higher dimensions. In passing, he was the first person to write down a metric for non-Euclidean geometry (more than a decade before Beltrami's publication of 1868, which legitimized non-Euclidean geometry). This lecture earned him the right to teach in a German university.

Riemann also did important work on shock waves, and shares with WEIERSTRASS [VI.44] the honor of introducing the methods of complex function theory in the study of MINIMAL SURFACES [III.96 §3.1], where he was led to several new solutions of the Plateau problem, which asks for the surface of least area spanning a given curve in space.

The distinguished complex analyst Lars Ahlfors once described Riemann's complex analysis as consisting of "almost cryptic messages to the future" and said that his mapping theorem was given in a form that "would defy any attempt at proof, even with modern methods," and it is true that Riemann's presentation is more visionary than precise. But his vision described a geometric setting for complex function theory that, as Ahlfors's own work indicates, remains fertile over 150 years after it was written.

Further Reading

- Laugwitz, D. 1999. *Bernhard Riemann, 1826–1866. Turning Points in the Conception of Mathematics*, translated by A. Shenitzer. Boston, MA/Basel: Birkhäuser.
- Riemann, G.F.B. 1990. *Gesammelte Werke, Collected Works*, edited by R. Narasimhan, third edn. Berlin: Springer.

Jeremy Gray

VI.50 Julius Wilhelm Richard Dedekind

b. Brunswick, Germany, 1831; d. Brunswick, Germany, 1916
Algebraic number theory; algebraic curves; set theory; foundations of mathematics

Dedekind spent most of his life as a professor at the Technische Hochschule in Brunswick, Germany (his and GAUSS's [VI.26] home town), having spent the years 1858–62 at the *Polytechnikum* in Zürich (which later

became known as ETH). He obtained his mathematical education at Göttingen, being Gauss's last Ph.D. student and subsequently a pupil of DIRICHLET [VI.36] and RIEMANN [VI.49]. Dedekind was a retiring man with, as KLEIN [VI.57] said, a "contemplative nature;" he remained a bachelor, living with his mother and sister. Nevertheless, he had an impact upon a select group of contemporaries (especially CANTOR [VI.54], FROBENIUS [VI.58], and Heinrich Weber) through his rich correspondence.

A key figure in the emergence of modern set-theoretic mathematics, and particularly the notion of a mathematical structure, Dedekind is best known for his work on the foundations of the REAL NUMBER SYSTEM [I.3 §1.4]. His main contribution, however, was in algebraic number theory. Indeed, he shaped modern number theory as we know it, presenting it as a theory of ideals in rings of integers (see ALGEBRAIC NUMBERS [IV.1 §§4–7]). This was first made public in 1871, within Supplement X to his edition of Dirichlet's *Vorlesungen über Zahlentheorie*, where he established unique decomposition of ideals into prime ideals for all rings of algebraic integers. In the process, he formulated the concepts of field, ring, ideal, and module (see [I.3 §2.2] and [III.83]), always within the particular context of the complex numbers. It was also in the context of algebra (Galois theory) and number theory that Dedekind started systematic work with quotient structures, isomorphisms, homomorphisms, and automorphisms.

In subsequent editions of Dirichlet's *Vorlesungen* (1879 and 1894) Dedekind went on refining his presentation of ideal theory, making it more purely set-theoretic. In 1882, together with Weber, he offered a theory of ideals in fields of algebraic functions, which made it possible to give a rigorous treatment of Riemann's results on algebraic curves up to the RIEMANN-ROCH THEOREM [V.34]. This work paved the way for modern algebraic geometry.

Intimately linked with Dedekind's work in algebra and number theory were his reflections on the foundations of the real number system. In 1858 (published 1872) he formulated a definition of the real numbers using what are now known as "Dedekind cuts" in the set of rational numbers. During the 1870s (published 1888) he elaborated a purely set-theoretic definition of the natural numbers as "simply infinite" sets, which led him to crystallize the DEDEKIND-PEANO AXIOMS [III.69]. In this work, as in his more advanced research,

sets, structures, and mappings form the essential building blocks, the very foundations of pure mathematics. In the light of (now superseded) conceptions of logic, this led Dedekind to the view that "arithmetic (algebra, analysis) is only a part of logic." From a modern viewpoint, his contributions show that SET THEORY [IV.22] is a sufficient foundation for classical mathematics. Thus he contributed as much as anybody else to the set-theoretic reformulation of modern mathematics.

Further Reading

- Corry, L. 2004. *Modern Algebra and the Rise of Mathematical Structures*, second revised edn. Basel: Birkhäuser.
 Ewald, W., ed. 1996. *From Kant to Hilbert: A Source Book in the Foundations of Mathematics*, two volumes. Oxford: Oxford University Press.
 Ferreirós, J. 1999. *Labyrinth of Thought. A History of Set Theory and Its Role in Modern Mathematics*. Basel: Birkhäuser.

José Ferreirós

VI.51 Émile Léonard Mathieu

b. Metz, France, 1835; d. Nancy, France, 1890

Student at the École Polytechnique; Docteur ès sciences with thesis on transitive functions (1859); Professor of Mathematics: Besançon (1869–74), Nancy (1874–90)

Mathieu is known for the functions that take his name, which he discovered while solving the two-dimensional wave equation for the vibrations of an elliptical membrane. These functions, which are special cases of the hypergeometric function, are particular solutions of *Mathieu's equation*:

$$\frac{d^2 u}{dz^2} + (a + 16q \cos 2z)u = 0,$$

where a and q are constants that depend on the physical problem.

Mathieu is also known for his discovery of the five Mathieu groups. These were the first SPORADIC SIMPLE GROUPS [V.8] (meaning that they did not fit into one of the known infinite families of simple groups) to be found. It is now known that there are twenty-six such groups altogether, although it was almost a century after Mathieu before a sixth one was found.

VI.52 Camille Jordan

b. Lyons, France, 1838; d. Milan, Italy, 1922

Nominally an engineer until 1885; teacher of mathematics, École Polytechnique and Collège de France (1873–1912)

Jordan was the leading group theorist of his generation. His immense *Traité des Substitutions et des Équations Algébriques* (1870), which brought together all his earlier results on PERMUTATION GROUPS [III.70] and provided a synthesis of GALOIS's [VI.41] ideas, remained a cornerstone for group theorists for many years. Included in the *Traité*, in the chapter on what he calls linear substitutions (now written in matrix form as $\mathbf{y} = A\mathbf{x}$), is the definition of what today is called the JORDAN NORMAL FORM [III.45] of a matrix, although in 1868 WEIERSTRASS [VI.44] had already defined an equivalent normal form.

Jordan is also known for his work in topology, especially for the theorem now known as the *Jordan curve theorem*. This states that a simple closed curve in the plane separates the plane into two disjoint regions, an inside and an outside, and it was given by him in his influential *Cours d'Analyse* (1887). Although the theorem appears obvious, the proof, as Jordan recognized, is difficult and the one he gave was incorrect. (The proof is relatively easy for smooth curves; the difficulties arise when dealing with nowhere-smooth curves, such as the Koch snowflake.) The first rigorous proof was given by Oswald Veblen in 1905. There is a stronger form of the theorem, known as the Jordan–Schönflies theorem, which states that in addition the two regions of the plane, the inside and the outside, are homeomorphic to the standard circle in the plane. Unlike the original theorem, this stronger form of the theorem cannot be generalized to higher dimensions, a famous counterexample being the Alexander horned sphere.

VI.53 Sophus Lie

b. Nordfjordeid (western Norway), 1842; d. Oslo, 1899

Transformation groups; Lie groups; partial differential equations

Lie was twenty-six when he discovered that, in his own words, he “harbored a mathematician.” Before then he had primarily wanted to be an observational astronomer. Later in life, looking back on his career, he said that it was the “audacity of his thinking” more than any formal knowledge and education that had given him a position among the foremost of mathematicians. During a career spanning more than thirty years, Lie produced almost eight thousand pages of mathematics, making him one of the most productive mathematicians of his time.

Lie graduated in general science from the university in Oslo in 1865 but without showing any special aptitude for mathematics. It was not until 1868, when he

attended a lecture by the Danish geometer Hieronymus Zeuthen on the work of Chasles, MÖBIUS [VI.30], and Plücker, that he became inspired by modern geometry. He studied the works of Poncelet (projective geometry) and Plücker (line geometry), and wrote a dissertation on “imaginary geometry,” that is, geometry based on complex numbers. In the fall of 1869 he traveled to Berlin, Göttingen, and Paris, where he met mathematicians who would remain friends and colleagues for the rest of his life. In Berlin he met KLEIN [VI.57], in Göttingen he met Clebsch, and in Paris, where he was joined by Klein, he met Darboux and JORDAN [VI.52]. These two had a particular influence on him—Darboux through his theory of surfaces and Jordan through his knowledge of group theory and the work of GALOIS [VI.41]—with the result that he (and Klein) began to recognize the value of group theory for the study of geometry. Lie and Klein published three joint papers on geometrical topics, including one on the so-called Lie line–sphere transformation (the contact transformation, which is a transformation that maps straight lines into spheres and principal tangent curves into curvature lines; and then the study of the geometrical entities that are invariant under such transformations).

When Klein prepared what was to become his famous “Erlanger Programm” (his characterization of geometry as properties invariant under a group action), Lie was with him. This work later created a deep rift between them. (Friendship turned into aloofness and hostility and culminated in the following statement by Lie in 1893: “I am no pupil of Klein’s, nor is the reverse the case, although this would be nearer to the truth.”)

Lie returned (after his first trip abroad) to Oslo and, in 1872, a chair of mathematics at the university was created especially for him. During the early 1870s Lie worked on turning his line–sphere transformation into a general theory of contact transformations. From 1873 he worked on a systematic study of continuous transformation groups (today known as LIE GROUPS [III.50 §1]), his aim being to classify LIE ALGEBRAS [III.50 §§2, 3] and apply the results to the solution of differential equations. He also published studies on MINIMAL SURFACES [III.96 §3.1]. In Norway, however, there was no scientific milieu, and he felt very isolated. In 1884 Klein and his friend Adolf Mayer in Leipzig tried to help him by sending their student Friedrich Engel to study with him and to help him with the formulation and writing of his new ideas. The work that Engel and Lie started together resulted in three volumes, *Theorie der Transformationsgruppen* (1888–93).

In 1886 Lie accepted the professorship in Leipzig (in succession to Klein, who had moved to Göttingen). In Leipzig he became a leading mathematician and a central figure in the European community of mathematicians. Promising new students from both France and the United States were sent to study with him. Besides teaching he continued his research on transformation groups and differential equations, and he solved the so-called Helmholtz space problem (characterizing the geometry of space in terms of groups of transformations). In 1898, the year before he died, Lie returned to Oslo to take up a position created especially for him.

The theory of transformation groups, which Lie initiated and developed in the study of differential equations, has grown into a field of its own, the theory of Lie groups and Lie algebras, which today permeates large parts of mathematics and mathematical physics.

Further Reading

- Borel, A. 2001. *Essays in the History of Lie Groups and Algebraic Groups*. Providence, RI: American Mathematical Society.
- Hawkins, T. 2000. *Emergence of the Theory of Lie Groups*. New York: Springer.
- Laudal, O. A., and B. Jahrien, eds. 1994. *Proceedings, Sophus Lie Memorial Conference*. Oslo: Scandinavian University Press.
- Stubhaug, A. 2002. *The Mathematician Sophus Lie*. Berlin: Springer.

Arild Stubhaug

VI.54 Georg Cantor

b. Saint Petersburg, Russia, 1845; d. Halle, Germany, 1918
Set theory; transfinite numbers; the continuum hypothesis

Although born in Russia, Cantor was raised and educated in Prussia and spent his entire career as professor of mathematics at the University of Halle. He studied at the Universities of Berlin and Göttingen with KRONECKER [VI.48], KUMMER [VI.40], and WEIERSTRASS [VI.44], and received his Ph.D. from the University of Berlin in 1867. His dissertation, “De aequationibus secundi gradus indeterminatis” (“On indeterminate equations of the second grade”), dealt with work in number theory on Diophantine equations, work that had been pioneered by LAGRANGE [VI.22], GAUSS [VI.26], and LEGENDRE [VI.24]. The following year he accepted a position in the mathematics department at the University of Halle, where he spent the rest

of his academic career. There, his *Habilitationsschrift* was again devoted to number theory, and dealt with transformations of ternary quadratic forms.

It was at Halle that Cantor's colleague, Eduard Heine, was working on difficult problems involving trigonometric series, and he interested Cantor in the problem of determining the conditions under which a trigonometric series of the form

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} (a_n \sin nx + b_n \cos nx)$$

uniquely represented a given function. In other words, could it be that two different trigonometric series could represent the same function? Heine had shown, in 1870, that if $f(x)$ is continuous in general (i.e., for all but a finite number of points of discontinuity, at which points Heine added that the function need not necessarily be finite), the representation is unique if we insist that the series is uniformly convergent to f in general. Cantor was able to establish much more general results, and in five papers written between 1870 and 1872 he was able to show that such representations were unique even if an infinite number of exceptional points were allowed, so long as these exceptional points (i.e., points at which the function failed to be continuous) were distributed over the domain of the function's definition in a particular way, constituting what Cantor called “point sets of the first species.” His studies of these and related point sets eventually led Cantor to his much more abstract and powerful theory of sets and transfinite numbers.

Point sets of the first species were sets P for which, given its sequence of *derived sets* (the derived set P' of a set P is the set of all the limit points of P), there was some finite n such that the n th derived set $P^{(n)}$ of P was finite, and thus the $(n + 1)$ st derived set was empty, i.e., $P^{(n+1)} = \emptyset$. It was Cantor's subsequent study of infinite linear point sets that would eventually lead to his creation of transfinite set theory in the 1880s. (For more details about this, see SET THEORY [IV.22 §2].)

Before he did so, Cantor first began to explore the implications of his work on trigonometric series and the structure of the real numbers in several papers, one of which was to revolutionize mathematics in a fundamental way. The first of these papers was published in 1874, and bore the innocuous title “Über eine Eigenschaft des Inbegriffes aller reellen algebraischen Zahlen” (“On a property of the collection of all real algebraic numbers”). In this paper, Cantor proved that the set of all algebraic real numbers was COUNTABLY INFINITE [III.11]. What was revolutionary about the paper,

however, was that he also proved that the set of all real numbers was *not* countable, and must be of a higher order of infinity than the countably infinite set of natural numbers. He returned to this result in 1891 with a new approach, the groundbreaking method of diagonalization, to prove in a very direct way that the set of real numbers is uncountably infinite. Cantor's second important paper of the decade appeared in 1878, "Ein Beitrag zur Mannigfaltigkeitslehre" ("A contribution to the theory of aggregates"), in which he proved (with a partly faulty argument) the invariance of dimension, a theorem first correctly proved by BROUWER [VI.75] in 1911.

Between 1879 and 1884 Cantor published six papers designed to outline the basic elements of his new thinking about sets. He first considered what happened if a set were not of the first species by introducing symbols for the infinite indices needed to identify such sets. For example, a set P was said to be of the *second* species if there was no finite n such that the n th derived set P^n of P was finite. He then considered the case in which the intersection of all the derived sets of P (namely $P', P'', \dots, P^n, \dots$) was again an infinite set, which he designated P^∞ . This set, since it was infinite, had a derived set as well, $P^{\infty+1}$, and this led in fact to an entire sequence of derived sets of the second species: $P^\infty, P^{\infty+1}, \dots, P^{\infty+n}, \dots, P^{2^\infty}, \dots$.

In his first papers on infinite linear sets, these indices for derived sets remained "infinitary symbols": that is, devices for distinguishing between different sets. But in his *Grundlagen einer allgemeinen Mannigfaltigkeitslehre* ("Foundations of a general set theory"), published in 1883, these symbols became the first transfinite numbers: the transfinite ordinal numbers. These numbers began with ω , the transfinite ordinal number representing the sequence of natural numbers $1, 2, 3, \dots$, which could also be thought of as the first infinite ordinal number after all of the finite whole numbers. In the *Grundlagen*, Cantor not only devised the basic features of a transfinite arithmetic for these numbers, but he provided a detailed philosophical defense of the new numbers. Acknowledging the revolutionary nature of what he was introducing, he argued that the new concepts were necessary in order to achieve precise mathematical results that he could obtain by no other means.

Cantor's best-known mathematical creation, however, the transfinite cardinal numbers, which he denoted using the Hebrew letter aleph, were introduced

only later, in the 1890s. They were first given full exposition in a pair of papers (1895, 1897) that constituted his "Beiträge zur Begründung der transfiniten Mengenlehre" ("Contributions to the founding of transfinite set theory"). In two articles published in *Mathematische Annalen*, he not only set out his theory of transfinite ordinal and cardinal numbers, as well as their arithmetics, but also explained his theory of order types, namely the different properties exhibited by the sets of natural, rational, and real numbers considered in their natural orders. There he also stated (but could not prove) his famous CONTINUUM HYPOTHESIS [IV.22 §5], namely that the power (or cardinal number) of the continuum of all real numbers \mathbb{R} is the next largest infinite set (or cardinal number) after the countably infinite set of natural numbers \mathbb{N} , the cardinality of which was taken to be \aleph_0 . Cantor expressed the continuum hypothesis algebraically as the statement that $2^{\aleph_0} = \aleph_1$.

By the end of his career Cantor had received honorary degrees from foreign universities and the Copley Medal of the Royal Society for his great contributions to mathematics, but there were problems with set theory that were beyond his capacities to remedy. The most disturbing for many mathematicians were the "antinomies" of set theory: the paradoxes put forward by the likes of Burali-Forti and RUSSELL [VI.71]. In 1897 the former published the paradox arising from the collection of *all* ordinal numbers, the ordinal number of which should be an ordinal number greater than any in the collection of all ordinal numbers. In 1901 the latter discovered the paradox of the class of all classes that are not members of themselves: is it a member of itself or not? (See THE CRISIS IN THE FOUNDATIONS OF MATHEMATICS [II.7].) Cantor himself was aware of the contradictions that arose from considering the collections of all transfinite ordinal or cardinal numbers, and what their ordinal or cardinal numbers might be. The solution Cantor adopted was to regard such collections as too large, and not really sets at all, but "inconsistent aggregates" as he called them. Others, like Zermelo, began to axiomatize set theory in an effort to exclude the possibility of contradictions. The two most powerful results of the twentieth century to complement Cantor's work are those of GÖDEL [VI.92] (who established the consistency of the continuum hypothesis with ZERMELO-FRAENKEL SET THEORY [IV.22 §3]) and Paul Cohen (who determined the independence of the continuum hypothesis from Zermelo-Fraenkel set theory), the latter finally establishing the impossibility of proving the continuum hypothesis.

Cantor's legacy for the history of mathematics has truly been revolutionary. Above all, his transfinite set theory for the first time gave mathematicians the means of dealing with concepts of the infinite in a careful and precise way.

Further Reading

Dauben, J. W. 1990. *Georg Cantor. His Mathematics and Philosophy of the Infinite*. Princeton, NJ: Princeton University Press. (First published in 1978 by Harvard University Press.)

———. 2005. Georg Cantor and the battle for transfinite set theory. In *Kenneth O. May Lectures of the Canadian Society for History and Philosophy of Mathematics*, edited by G. Van Brummelen and M. Kinyon, pp. 221–41. New York: Springer.

———. 2005. Georg Cantor. Paper on the “Foundations of a general set theory” (1883). In *Landmark Writings in Western Mathematics 1640–1940*, edited by I. Grattan-Guinness, pp. 600–12. London: Routledge.

Tapp, C. 2005. *Kardinalität und Kardinäle. Wissenschaftshistorische Aufarbeitung der Korrespondenz zwischen Georg Cantor und katholischen Theologen seiner Zeit*. Stuttgart: Franz Steiner.

Joseph W. Dauben

VI.55 William Kingdon Clifford

b. Exeter, England, 1845; d. Madeira, Portugal, 1879

Geometry; complex function theory; popularization of mathematics

Clifford went up to Trinity College Cambridge in 1863. He graduated from there in 1867 as 2nd Wrangler and also came second in the more demanding Smith's prize examination. In 1868 he became a Fellow of Trinity, leaving in 1871 to become the professor of applied mathematics at University College London. He died of tuberculosis in 1879.

A versatile mathematician, regarded by many as the best of his generation, Clifford's favorite field was geometry, over which he ranged widely, proving new results in classical Euclidean geometry as well as in projective and differential geometry. He was the first English mathematician to appreciate the work of RIEMANN [VI.49] on differential geometry, and published a translation of Riemann's paper “On the hypotheses that lie at the foundations of geometry” in 1873. He endorsed Riemann's fundamental reformulation of geometry, and went even further in speculating that the curvature of physical space might explain the motion of matter. He also made a significant application of

the RIEMANN-ROCH THEOREM [V.34], and was among the first to analyze the complicated topological nature of a RIEMANN SURFACE [III.81] by showing how to dissect any Riemann surface into simple pieces in a standard way. He was the first to study a geometry locally equivalent to plane geometry but topologically distinct (the flat torus, also known today as the *Clifford-Klein space form* after KLEIN's [VI.57] later more detailed study of it). In algebra, he invented the biquaternions (these are like quaternions, but have complex numbers as coefficients).

Clifford was regarded as a marvelous lecturer until his health broke, and he was a successful popularizer and essay writer. He forcefully adopted the view that geometry was a matter of experience, not a priori truth. He was a friend of T. H. Huxley and was sympathetic to humanism in philosophy.

Further Reading

Clifford, W. K. 1968. *Mathematical Papers*, edited by R. Tucker. New York: Chelsea. (First published in 1882.)

Jeremy Gray

VI.56 Gottlob Frege

b. Wismar, Germany, 1848; d. Bad Kleinen, Germany, 1925

Logic; foundations of mathematics; paradox

Frege was a precursor of modern logic, in that many of the hallmarks of contemporary logic appear first in his writing. His work has also been singularly influential outside the foundations of mathematics, especially in the philosophy of language.

Frege was trained at Jena and Göttingen, receiving a Ph.D. under Ernst Schering in 1873. His Ph.D. thesis addressed the spatial representation of imaginary elements in geometry, and his 1874 Habilitation essay at Jena worked out some basic details of what we would now call “iteration theory.” Though his early work gave no obvious sign of the revolutionary work to come, with hindsight one can discern a foundational motif running through even the apparently conventional mathematics of the early work: a conviction that arithmetic was in some way or other logical, and that geometry was fundamentally different and less general because it was grounded in spatial intuition. This is an especially salient concern in some of his areas of early research, such as Plücker's line geometry and RIEMANN's [VI.49]

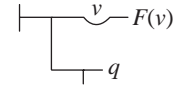
complex analysis, where the role of visual representation was a matter of some dispute. Frege sought to resolve the dispute by deriving arithmetic and analysis rigorously from logical principles. His motivation was not so much a desire for certainty: rather, he held that only “gap-free” proofs can reveal a science’s fundamental principles.

Among the features of contemporary logic appearing first in Frege’s core logical writings (*Begriffsschrift* (1879) and *Grundgesetze der Arithmetik* (volume 1, 1893; volume 2, 1903)) are the following.

- (i) Inferences are analyzed within a quantified logic of propositions, which extends to relations as well as to propositions of subject–predicate form. We would today describe Frege’s logical system as a higher-order predicate calculus.
- (ii) Forms from syllogistic logic (such as “All As are Bs”) are interpreted as quantified conditionals (“For all x , if x is an A, then x is a B”), in the way that is now so standard as to seem inescapable, presenting implicitly the point that the underlying logical form of a proposition may differ from its surface grammar.
- (iii) The syntax of the language is explicitly displayed, and inferences are carried out strictly in accordance with the form of statements by explicitly stated rules.
- (iv) Rules of inference and axioms are distinguished; the consequence relation and conditionals are distinguished.
- (v) “Function” is taken as an undefined primitive concept. (This was a contentious move. Some mathematicians of the time, including one of Frege’s teachers, Alfred Clebsch, held the concept of function to be too vague to serve as a basic building block.) A sharp distinction is enforced between functions and the things (called objects) that can be arguments of functions.
- (vi) Quantifiers can be iterated, making possible the logical representations of distinctions such as that between uniform and pointwise convergence.

However, any simple catalogue of novelties understates the crystalline sharpness of Frege’s logical writing when compared with works with similar aims, such as the later *Principia Mathematica* of Whitehead and RUSSELL [VI.71]. It would be several decades before logicians approached Frege’s standards for exactness and clarity. The notation, however, seemed unwieldy

to readers at the time (and since). Here, for example, is the statement “if not q , then every v is F ” ($\neg q \Rightarrow (\forall v)F(v)$) in Frege’s notation:



(Here \neg represents negation, \forall is the universal quantifier, and the long vertical line represents the conditional.)

Frege also wrote an informal treatise, *Grundlagen der Arithmetik* (1884), which has had a profound influence on English-language philosophy since its translation in 1950. Its account of number contains the first hint of the tension that would collapse the project from within. Frege sets out conditions that a definition of number must satisfy to be counted as “acceptable.” However, when formalized these lead to a contradiction, of a similar type to Russell’s paradox (on the set of all sets that are not members of themselves). This problem escaped Frege’s notice until Russell alerted him to it in a letter of 1903. Frege’s reaction (“arithmetic totters”) has been taken to be an overreaction to the failure of one set of axioms among many possible ones. But in Frege’s view, the problem was not with the specific axioms, but rather that any logically adequate weakening appeared to violate some principle he took to be grounded in the nature of thought. Recently, many logicians who do not share Frege’s often baroque-seeming metaphysics of concepts have shown that some natural consistent weakenings of Frege’s system do support the derivation of the mathematics Frege aimed to reconstruct.

The years after 1903 brought personal tragedies in Frege’s life and he ceased serious work for over a decade. Though he resumed writing in 1918 with a series of philosophical articles, his only research in mathematics was a brief jotted effort to found arithmetic on geometry, rather than logic, indicating his conclusion that his logical program had failed.

Further Reading

A particularly detailed recent example of “neo-Fregean” reconstructions of Frege’s foundations of arithmetic appears in John Burgess’s *Fixing Frege* (Princeton University Press, Princeton, NJ, 2005). Many of the classic papers on the technical details of reconstructing Frege’s philosophy of logic are reprinted in *Frege’s Philosophy of Mathematics* (Harvard University Press, Harvard, MA, 1995) by William Demopoulos.

Jamie Tappenden

VI.57 Christian Felix Klein

b. Düsseldorf, Germany, 1849; d. Göttingen, Germany, 1925
Higher geometry; function theory; theory of algebraic equations; pedagogy

Klein had originally intended to be a physicist but during the course of his studies with Julius Plücker in Bonn, with whom he studied both mathematics and physics, he turned to mathematics, receiving his doctorate for a thesis on line geometry in 1868. After Plücker's death in 1868 he went to Göttingen to study with Alfred Clebsch, where he worked exclusively on mathematics. In 1869–70 he spent some months in Berlin studying with WEIERSTRASS [VI.44] and KUMMER [VI.40] before joining LIE [VI.53] for a trip to Paris to see HERMITE [VI.47]. After passing his habilitation in Göttingen in 1871, he took positions successively at Erlangen, Munich, and Leipzig, returning to Göttingen in 1886, where he remained until he retired (because of poor health) in 1913. In 1875 he married Anna Hegel, a granddaughter of the philosopher Georg Wilhelm Friedrich Hegel.

In 1872 Klein published his celebrated “Erlanger Programm,” a creative and unified conception of geometry. Building on a paper of CAYLEY [VI.46] of 1859 in which Cayley had shown how to deduce EUCLIDEAN GEOMETRY [I.3 §6.2] from PROJECTIVE GEOMETRY [I.3 §6.7], Klein applied his knowledge of group theory (learned from JORDAN [VI.52] in Paris) to create a hierarchy of all geometries. He had recognized that each geometry could be characterized by a group of transformations and classified accordingly (see SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3 §6.1]). The classification showed, as Klein had anticipated, that of all the geometries, projective is the most basic and that the others, e.g., affine, hyperbolic, Euclidean, etc., are subsumed at some level beneath it. Moreover, it was clear from his construction that a contradiction in NON-EUCLIDEAN GEOMETRY [II.2 §§6–10] would simultaneously involve a contradiction in Euclidean geometry.

Klein regarded his work in function theory as his greatest achievement. As his career progressed, he moved more and more from Plücker's and Clebsch's strictly geometric viewpoint toward the wider outlook embraced by RIEMANN [VI.49], who had regarded analytic functions as given by conformal mappings between given domains. In his “Riemanns Theorie der algebraischen Funktionen und ihrer Integrale” (1882), Klein provided a geometric treatment of function

theory in which he fused Riemann's ideas with the rigorous power-series methods of Weierstrass.

In 1882, when he was at the height of his powers, Klein's health broke down. His attempt to keep up with POINCARÉ [VI.61] in the race to develop the theory of automorphic functions (which are generalizations of periodic functions such as trigonometric functions, ELLIPTIC FUNCTIONS [V.34], etc.), during which he had proved his famous Grenzkreis (boundary circle) Theorem, had left him exhausted, and he was never again able to work with such intensity and at such a high level.

After his breakdown Klein's interest shifted progressively from research toward pedagogy. In his efforts to modernize mathematical education he developed outstanding organizational skills and initiated important and far-reaching editorial projects ranging from the preparation of lecture notes to coediting the twenty-four-volume *Encyklopädie der mathematischen Wissenschaften* (1896–1935). He was an editor of the *Mathematische Annalen* for almost fifty years, and was among the founding members of the Deutsche Mathematiker-Vereinigung (1890). He also played an active role in establishing mathematical applications in science and engineering, as well as promoting the better understanding of mathematics by engineers.

Among Klein's other achievements were important results in the theory of algebraic equations (through a consideration of the icosahedron he obtained a complete theory of the general fifth-degree equation (1884)) and in mechanics, in which, jointly with Arnold Sommerfeld, he developed the theory of the gyroscope (1897–1910). He also worked on ideas involving the application of group theory to the theory of relativity, producing papers on the LORENTZ GROUP [IV.13 §1] (1910) and gravitation (1918). Klein was an international figure who traveled widely, including to the United States and the United Kingdom, and he played a significant role in the first International Congresses of Mathematicians. His many foreign students included several from the United States, e.g., Maxime Bôcher and William Fogg Osgood, and a number of women, notably Grace Chisholm Young and Mary Winston.

Klein's achievements made Göttingen the scientific center of Germany and one of the mathematical centers of the world. He possessed an outstanding ability to “see” the truth in mathematical statements and to bring mathematical fields together without feeling the necessity for detailed calculations and justification (which he left to his students and others). He believed strongly in the unity of mathematics.

Further Reading

- Frei, G. 1984. Felix Klein (1849–1925), a biographical sketch. In *Jahrbuch Überblicke Mathematik*, pp. 229–54. Mannheim: Bibliographisches Institut.
- Klein, F. 1921–23. *Gesammelte mathematische Abhandlungen*, three volumes. Berlin: Springer. (Reprinted, 1973. Volume 3 contains lists of Klein’s publications, lectures, and dissertations directed by him.)
- . 1979. *Development of Mathematics in the 19th Century*, translated by M. Ackerman. Brookline, MA: MathSciPress.

Rüdiger Thiele

VI.58 Ferdinand Georg Frobenius

b. Berlin, 1849; d. Berlin, 1917

Analysis; linear algebra; number theory; theory of groups; character theory

After school in Berlin, Frobenius (who suppressed his first name and wrote mainly as G. Frobenius) spent one semester studying mathematics and physics in Göttingen, then returned to Berlin where he studied under KRONECKER [VI.48], KUMMER [VI.40], WEIERSTRASS [VI.44], and others. He wrote his doctoral dissertation (in Latin) supervised by Weierstrass, in 1870, on infinite series representations of analytic functions of one variable. For four years he worked as a schoolteacher in Berlin before he became Außerordentlicher Professor (associate professor) at Berlin University. After less than two years, in 1875, he was called to a full professorship at the Eidgenössische Technische Hochschule in Zürich, where he remained until 1892, when he returned to Berlin as successor to KRONECKER [VI.48]. He retired in 1916, and died one year later.

His early contributions were to analysis and the theory of differential equations. Later he wrote mainly on theta functions, algebra, and number theory. One of his well-known contributions lies across group theory and number theory. Given a polynomial with coefficients in an algebraic number field, one may ask for the degrees of the irreducible factors that occur when it is reduced modulo a prime ideal. In particular, one may ask for the “density” (suitably defined) of the set of prime ideals modulo which a given pattern of irreducible-factor degrees arises. Pursuing ideas of Kronecker, Frobenius proved that, if the GALOIS GROUP [V.24] is the SYMMETRIC GROUP [III.70], then that density is the proportion of elements of the group whose

cycle structure is the pattern of degrees. He conjectured that this should be true whatever the Galois group. A tool he used for this led to the name “Frobenius automorphism” for the natural generator $a \mapsto a^q$ of the Galois group of a finite extension of the field \mathbb{F}_q . The conjecture was proved by N. G. Chebotaryov in 1925 and is now known as the Chebotaryov density theorem, or, sometimes, the Frobenius–Chebotaryov density theorem.

Another well-known and important contribution was to the theory of matrices and linear transformations, where Frobenius introduced the minimal polynomial and other invariants (the elementary divisors).

Frobenius is best known for his work in finite group theory. Like Otto Hölder and WILLIAM BURNSIDE [VI.60], he focused for a time on the search for FINITE SIMPLE GROUPS [V.8]. His greatest contribution, however, is his invention of the theory of GROUP CHARACTERS [IV.9]. This emerged unexpectedly in 1896 out of his study of group determinants. These are the determinants of square matrices with rows and columns indexed by the members of a finite group G , and with (a, b) -entry $x_{ab^{-1}}$, where the x_g are independent variables, one for each element g of G . His interest, stimulated by correspondence with DEDEKIND [VI.50], was in how the group determinant factorizes as a polynomial in these variables. This problem led Frobenius to the discovery of certain sets of complex numbers, which he called *Gruppencharactere*, one for each conjugacy class in the group, that arose as the solutions of sets of linear equations connected with the group. Nowadays they are defined differently: for each complex linear representation ρ of the group G (that is, homomorphism $\rho : G \rightarrow \text{GL}_n(\mathbb{C})$, where $\text{GL}_n(\mathbb{C})$ is the group of $n \times n$ invertible matrices over \mathbb{C}), the associated character χ is the map $G \rightarrow \mathbb{C}$ such that $\chi(g) = \text{trace } \rho(g)$ for $g \in G$. Frobenius proved the orthogonality relations, recognized the connection of his characters with matrix representations of the group, calculated the character tables of the symmetric groups, the alternating groups, and the Mathieu groups, and used properties of induced characters to prove his famous theorem that a transitive permutation group in which no element other than the identity fixes two or more points has a regular normal subgroup (that is, a subgroup consisting of the identity together with the fixed-point-free elements of the group). To this day no purely group-theoretic proof of this theorem has been found. In recognition of his contribution such groups are now called Frobenius groups. Through character theory and representation theory, as developed

PUP: this sentence is fine as it is.

PUP: because Neumann’s articles are allowed to keep UK style, I have not changed the second hyphen here to an en-rule. OK?

by Frobenius for finite groups (and by his pupil, friend, and colleague Issai Schur for classical matrix groups), group theory found important applications in physics and chemistry a generation later.

Further Reading

- Begehr, H., ed. 1998. *Mathematik in Berlin: Geschichte und Dokumentation*, two volumes. Aachen: Shaker.
- Curtis, C. W. 1999. *Pioneers of Representation Theory: Frobenius, Burnside, Schur, and Brauer*. Providence, RI: American Mathematical Society.
- Serre, J.-P., ed. 1968. *F. G. Frobenius: Gesammelte Abhandlungen*, three volumes. Berlin: Springer.

Peter M. Neumann

VI.59 Sofya (Sonya) Kovalevskaya

b. Moscow, 1850; d. Stockholm, 1891
Partial differential equations; Abelian integrals

Kovalevskaya showed talent for mathematics at an early age but as a woman in mid-nineteenth-century Russia she was denied access to university. Unable to leave the country unescorted she married and in 1869 traveled to Heidelberg, where she was taught mathematics by Du Bois-Reymond. The following year she moved to Berlin to work with WEIERSTRASS [VI.44]. Berlin University was closed to women but Weierstrass agreed to tutor her privately. Under his supervision Kovalevskaya completed dissertations on partial differential equations (PDEs), Abelian integrals, and Saturn's rings, and in 1874 she became the first woman to receive a doctorate in mathematics. The dissertation on PDEs, which excited particular attention, contained the result now known as the CAUCHY-KOVALEVSKAYA THEOREM [IV.12 §§2.2, 2.4], an important tool in establishing the existence of analytic solutions of PDEs.

That same year Kovalevskaya returned to Russia and, unable to find a suitable position, temporarily abandoned mathematics. In 1880, at the invitation of CHEBYSHEV [VI.45], she gave a paper on Abelian integrals at a conference in Saint Petersburg. It was enthusiastically received and in 1881 she returned to Berlin. She saw Weierstrass frequently and devoted herself to the study of the propagation of light in a crystalline medium—a subject to which she had been led by studying the work of the French physicist Gabriel Lamé—and to the study of the rotation of a solid body about a fixed point. Later that year she moved to Paris to work with mathematicians there.

In 1883, championed by Mittag-Leffler, Kovalevskaya was appointed as a Privatdozent at the University of Stockholm. She also became an editor of *Acta Mathematica*, making her the first woman to join the board of a scientific journal. On behalf of *Acta* she liaised with mathematicians from Paris, Berlin, and Russia, providing an important link between Russian mathematicians and their western European counterparts. She continued to work on the rotation problem and in 1885 made the breakthrough that, three years later, would win her the prestigious *Prix Bordin* of the French Academy of Sciences. Prior to her work the problem had been completely solved for only two cases, both symmetrical. In the first, solved by EULER [VI.19], the center of gravity of the moving body coincides with the fixed point; and in the second, solved by LAGRANGE [VI.22], the center of gravity and the fixed point lie on the same axis. Kovalevskaya discovered that there was a third case, one that was asymmetrical and more complicated than the other two, which could also be solved completely. (It was later shown that there are no others.) The novelty of her results lay in her application of the recently developed theory of theta functions—the simplest elements from which ELLIPTIC FUNCTIONS [V.34] can be constructed—to solve Abelian integrals.

Kovalevskaya became a full professor of mathematics at the University of Stockholm in 1889, the first woman anywhere to achieve such a position. Shortly afterward, she was nominated by Chebyshev for corresponding membership of the Russian Academy of Sciences, her subsequent election breaking the gender barrier once again.

Further Reading

- Cooke, R. 1984. *The Mathematics of Sonya Kovalevskaya*. New York: Springer.
- Koblitz, A. H. 1983. *A Convergence of Lives. Sofia Kovalevskaya: Scientist, Writer, Revolutionary*. Boston, MA: Birkhäuser.

VI.60 William Burnside

b. London, 1852; d. West Wickham, England, 1927
Theory of groups; character theory; representation theory

Burnside's mathematical abilities first showed themselves at school. From there he won a place at Cambridge, where he read for the Mathematical Tripos and graduated as 2nd Wrangler in 1875. For ten years he remained in Cambridge as a Fellow of Pembroke College, coaching student rowers and mathematicians. In

1885, having published three very short papers, he was appointed professor at the Royal Naval College, Greenwich. He married in 1886 and the next year, at the age of thirty-five, he embarked on his career as a productive mathematician. He was elected as a Fellow of the Royal Society in 1893 on the basis of his contributions in applied mathematics (statistical mechanics and hydrodynamics), geometry, and the theory of functions. Although he continued to contribute to these areas throughout his working life, and added probability theory to his fields of interest during World War I, he turned to the theory of groups in 1893, and it is for his discoveries in this subject that he is remembered.

Burnside treated every aspect of the theory of finite groups. He was much concerned with the search for finite simple groups, and made the famous conjecture, finally proved by Walter Feit and John Thompson in 1962, that there are no simple groups of odd composite order (see THE CLASSIFICATION OF FINITE SIMPLE GROUPS [V.8]). He helped to develop character theory, which had been created by FROBENIUS [VI.58] in 1896, into a tool for proving theorems of pure group theory, using it in 1904 to spectacular effect when he proved his so-called $p^\alpha q^\beta$ -theorem: the theorem that groups whose orders are divisible by at most two different prime numbers are soluble. By asking, in effect, whether a group all of whose elements have finite order and which is generated by finitely many elements must be finite, he launched the huge area of research which for much of the twentieth century was known as the Burnside problem (see GEOMETRIC AND COMBINATORIAL GROUP THEORY [IV.10 §5.1]).

Although CAYLEY [VI.46] and the Reverend T. P. Kirkman had written about groups before him, he was the only British mathematician to work in group theory until Philip Hall started his mathematical career in 1928. Burnside's influential book *Theory of Groups of Finite Order* (1897) was written in the hope of "arousing interest among English mathematicians in a branch of pure mathematics which becomes the more interesting the more it is studied." Its influence in his own country was minimal, however, until several years after his death. It went to a second edition in 1911 (reprinted 1955), which differs from the first in that it has been substantially revised and, in particular, it includes chapters about the character theory of finite groups and its applications—mathematics which had been much developed by Frobenius, Burnside, and Schur over the fifteen years following the invention of character theory in 1896.

Further Reading

- Curtis, C. W. 1999. *Pioneers of Representation Theory: Frobenius, Burnside, Schur, and Brauer*. Providence, RI: American Mathematical Society.
- Neumann, P. M., A.J.S. Mann, and J. C. Thompson. 2004. *The Collected Papers of William Burnside*, two volumes. Oxford: Oxford University Press.

Peter M. Neumann

VI.61 Jules Henri Poincaré



b. Nancy, France, 1854; d. Paris, 1912

Function theory; geometry; topology; celestial mechanics; mathematical physics; foundations of science

Educated at the École Polytechnique and the École des Mines in Paris, Poincaré began his teaching career at the University of Caen in 1879. In 1881 he took up an appointment at the University of Paris where, from 1886, he held successive chairs until his death in 1912. He was of a retiring nature and did not attract graduate students, but his lecture courses provided the basis for a number of treatises, mostly in mathematical physics.

Poincaré came to international prominence in the early 1880s when, fusing ideas from complex function theory, group theory, non-Euclidean geometry, and the theory of ordinary linear differential equations, he identified an important class of automorphic functions. Named Fuchsian functions, in honor of the mathematician Lazarus Fuchs, they are defined on a disk and

remain invariant under certain discrete groups of transformations. Soon after, he identified the related but more complicated Kleinian functions, which are automorphic functions without a limit circle. His theory of automorphic functions was the first significant application of non-Euclidean geometry. It led to his discovery of the disk model of the hyperbolic plane and later inspired the UNIFORMIZATION THEOREM [V.37].

During the same period Poincaré began pioneering work on the qualitative theory of differential equations, motivated in part by an interest in some of the fundamental questions of mechanics, notably the problem of the stability of the solar system. What was new and important was his idea of thinking of the solutions in terms of curves rather than functions, i.e., thinking geometrically rather than algebraically, and it was this that marked a departure from the work of his predecessors, whose research had been dominated by power-series methods. From the mid 1880s he began applying his geometric theory to problems in celestial mechanics. His memoir on THE THREE-BODY PROBLEM [V.36] (1890) is famous both for providing the basis for his acclaimed treatise, *Les Méthodes Nouvelles de la Mécanique Céleste* (1892–99), and for containing the first mathematical description of CHAOTIC BEHAVIOR [IV.14 §1.5] in a dynamical system. Stability was also at the heart of his investigation into the forms of rotating fluid masses (1885). This work, which contained the discovery of new, pear-shaped figures of equilibrium, aroused considerable attention because of its important implications for cosmogony in relation to the evolution of binary stars and other celestial bodies.

Poincaré's work on Fuchsian functions and on the qualitative theory of differential equations led him to recognize the importance of the topology (or, as it was then called, *analysis situs*) of MANIFOLDS [I.3 §6.9]. And in the 1890s he began to study the topology of manifolds as a subject in its own right, effectively creating the powerful independent field of ALGEBRAIC TOPOLOGY [IV.6]. In a series of memoirs published between 1892 and 1904, the last of which contains the hypothesis known today as THE POINCARÉ CONJECTURE [IV.7 §2.4], he introduced a number of new ideas and concepts, including Betti numbers, THE FUNDAMENTAL GROUP [IV.6 §2], HOMOLOGY [IV.6 §4], and torsion.

A deep interest in physical problems lay behind Poincaré's achievements in mathematical physics. His work in potential theory forms a bridge between that of Carl Neumann on boundary-value problems and that of FREDHOLM [VI.66] on integral equations. He introduced

the "méthode de balayage" ("sweeping-out method") for establishing the existence of solutions to the DIRICHLET PROBLEM [IV.12 §1] (1890); and he had the idea that the Dirichlet problem itself should give rise to a sequence of EIGENVALUES AND EIGENFUNCTIONS [I.3 §4.3] (1898). In developing the theory for functions of several variables he was led to the discovery of new results in complex function theory. In *Électricité et Optique* (1890, revised 1901), which derived from his university lectures, he gave an authoritative account of the electromagnetic theories of Maxwell, Helmholtz, and Hertz. In 1905 he responded to Lorentz's new theory of the electron, coming close to anticipating Einstein's theory of SPECIAL RELATIVITY [IV.13 §1], thereby provoking controversy among later writers about the question of priority. And in 1911 he attended the first Solvay Conference on quantum theory, publishing an influential memoir (1912) in its favor.

As Poincaré's career developed, so too did his interest in the philosophy of mathematics and science. His ideas became widely known through four books of essays: *La Science et l'Hypothèse* (1902), *La Valeur de la Science* (1905), *Science et Méthode* (1908), and *Dernières Pensées* (1913). As a philosopher of geometry he was a proponent of the view, known as conventionalism, that it is not an objective question which model of geometry best fits physical space but is rather a matter of which model we find most convenient. By contrast, his position on arithmetic was intuitionist. On the question of foundational issues, he was largely critical. Although sympathetic to the goals of set theory, he attacked what he perceived as its counterintuitive results (see THE CRISIS IN THE FOUNDATIONS OF MATHEMATICS [II.7 §2.2]).

Poincaré's visionary geometric style led him to new and brilliant ideas, which frequently connected different branches of mathematics, but lack of detail often made his work hard to follow. At times his approach was censured for imprecision; it was in marked contrast to that of HILBERT [VI.63], his German counterpart, whose work was rooted in algebra and rigor.

Further Reading

- Barrow-Green, J. E. 1997. *Poincaré and the Three Body Problem*. Providence, RI: American Mathematical Society.
 Poincaré, J. H. 1915–56. *Collected Works: Œuvres de Henri Poincaré*, eleven volumes. Paris: Gauthier Villars.

VI.62 Giuseppe Peano

b. Spinetta, Italy, 1858; d. Turin, 1932

Analysis; mathematical logic; foundations of mathematics

Known above all for his (and DEDEKIND's [VI.50]) axiom system for the natural numbers, Peano made important contributions to analysis, logic, and the axiomatization of mathematics. He was born in Spinetta (Piedmont, Italy) as the son of a peasant, and from 1876 studied at the University of Turin, taking his doctoral degree in 1880. He remained there until his death in 1932, becoming full professor in 1895.

During the 1880s Peano worked in analysis, achieving what are generally considered to be his most important results. Particularly noteworthy are the continuous space-filling *Peano curve* (1890), the notion of *content* (a precedent of MEASURE THEORY [III.57]) developed independently by JORDAN [VI.52], and his theorems on the existence of solutions for differential equations of the first order (1886, 1890). The textbook he published in 1884, *Calcolo Differenziale e Principii di Calcolo Integrale*, partly based on lectures by his teacher Angelo Genocchi, was noteworthy for its rigor and critical style, and is counted among the very best nineteenth-century treatises.

The years 1889–1908 saw Peano dedicating himself intensively to symbolic logic, axiomatization, and producing the encyclopedic *Formulaire de Mathématiques* (1895–1908, five volumes). This ambitious assembly of mathematical results, compactly presented in the symbols of mathematical logic, was given completely without proofs. This was by no means standard at the time, but it shows what Peano expected from logic: it was supposed to bring precision of language and brevity, but not a greater level of rigor (something that was, by contrast, crucial for FREGE [VI.56]). In 1891, together with some colleagues, he founded the journal *Rivista di Matematica*, gathering around him an important group of followers.

Peano was an accessible man, and the way he mingled with students was regarded as “scandalous” in Turin. He was a socialist in politics, and a tolerant universalist in all matters of life and culture. In the late 1890s Peano became increasingly interested in elaborating a universal spoken language, “*Latino sine flexione*”; the last edition of the *Formulario* (1905–8) appeared in this language.

Peano followed closely the work of German mathematicians such as Hermann Grassmann, Ernst Schrö-

der, and Richard Dedekind; for example, the 1884 textbook defined the real numbers by Dedekind cuts, and in 1888 he published *Calcolo Geometrico Secondo l'Ausdehnungslehre di H. Grassmann*. In 1889 there appeared (notably in Latin) a first version of the famous PEANO AXIOMS [III.69] for the set of natural numbers, which he refined in volume 2 of the *Formulaire* (1898). It aimed at filling the most significant gap in the foundations of mathematics at a time when the *arithmetization* of analysis had essentially been completed. It is no coincidence that other mathematicians (Frege, Charles S. Peirce, and Dedekind) published similar work in the same decade. Peano's attempt is better rounded than Peirce's, but simpler and framed in more familiar terms than those of Frege and Dedekind; because of this, it has been more popular.

Peano's work on the natural numbers was at the crossroads of his diverse mathematical contributions, linking naturally his previous research in analysis with his later work on logical foundations, and being a necessary prerequisite for the *Formulaire* project. Actually, *Arithmetices Principia* can be regarded as a simplification, refinement, and translation into logical language (the “*nova methodo*” in its title) of Grassmann's *Lehrbuch der Arithmetik* (1861). Grassmann had striven to elaborate a stern deductive structure, stressing proofs by mathematical induction and recursive definitions. But curiously, unlike Peano, he did not postulate an axiom of induction; thus, Peano presented the basic assumptions much more clearly, bringing induction to center stage as the key defining property of the natural numbers.

Further Reading

- Borga, M., P. Freguglia, and D. Palladino. 1985. *I Contributi Fondazionali della Scuola di Peano*. Milan: Franco Angeli.
- Ferreirós, J. 2005. Richard Dedekind (1888) and Giuseppe Peano (1889), booklets on the foundations of arithmetic. In *Landmark Writings in Western Mathematics 1640–1940*, edited by I. Grattan-Guinness, pp. 613–26. Amsterdam: Elsevier.
- Peano, G. 1973. *Selected Works of Giuseppe Peano*, with a biographical sketch and bibliography by H. C. Kennedy. Toronto: University of Toronto Press.

José Ferreirós

VI.63 David Hilbert



b. Königsberg, Germany, 1862; d. Göttingen, Germany, 1943
Invariant theory; number theory; geometry;
International Congress of Mathematicians; axiomatics

HERMANN WEYL [VI.80] described his teacher Hilbert's style: "It is as if you were on a swift walk through a sunny open landscape; you look freely around, demarcation lines and connecting roads are pointed out to you, before you must brace yourself to climb the hill; then the path goes straight up...." Several themes balance in Hilbert's career as a mathematician. He wanted clarity, rigor, simplicity, and depth. Though he loved mathematics for its beauty, a beauty that transcends human failures, Hilbert saw mathematics as a social collaboration. A turning point came when he met MINKOWSKI [VI.64] and Adolf Hurwitz at university in Königsberg.

Hilbert wrote: "On unending walks we engrossed ourselves in the actual problems of the mathematics of the time; exchanged our newly acquired understandings, our thoughts and scientific plans; and formed a friendship for life." Later Hilbert became professor at Göttingen and, with KLEIN [VI.57], drew mathematicians from all over the world and turned that small city into a crossroads for mathematics—until Hitler destroyed it.

When he was a new Privatdozent, Hilbert decided he would study mathematics as he taught, and he resolved never to repeat lectures. He and Hurwitz decided to embark on a "systematic exploration" of mathematics, and he followed this pattern for the rest of his life. Hilbert's career divides easily into six periods: (i) algebra and algebraic invariants (1885–93); (ii) algebraic

number theory (1893–98); (iii) geometry (1898–1902); (iv) analysis (1902–12); (v) mathematical physics (1910–22); and (vi) foundations (1918–30). Remarkably, there is very little overlap. When Hilbert finished a subject, he was finished with it.

Hilbert's first breakthrough came in 1888 when he solved Gordan's problem, named after Paul Gordan, in a single bold move. Given a polynomial equation with at least two variables, some things about the polynomial change and some do not when you change coordinate systems. For example, with the real polynomial equation

$$ax^2 + bxy + cy^2 + d = 0,$$

if you rotate the coordinate system the equation changes dramatically, but the graph does not, and neither does the discriminant $b^2 - 4ac$. The discriminant is one invariant. In the general case—a more complicated class of polynomials and coordinate changes—there can be many invariants. Mathematicians suspected that a finite number of essentially different invariants existed for any given type of polynomial and class of coordinate changes. Was this so? Many mathematicians calculated individual examples industriously. Instead, Hilbert reasoned indirectly: what if there is no finite basis for a specific class of polynomials and transformations? He found that it was always possible to produce a contradiction. He concluded that there must be such a basis. At first this result was greeted with disbelief because he did not display a basis. Gordan said, "Das ist nicht Mathematik. Das ist Theology." However, the result was so powerful that it has been said that it killed algebraic invariant theory.

In 1893 Hilbert and Minkowski were asked by the German Mathematical Society to write a report on number theory. Hilbert chose ALGEBRAIC NUMBER THEORY [IV.1] and transformed the results of the nineteenth century into the study of algebraic NUMBER FIELDS [III.65]. The deep organizing structure Hilbert found eventually led to what has been called "the magnificent edifice of class field theory" (described in [V.31]).

Hilbert's classic *Foundations of Geometry*, first published in 1899 and revised many times, starts with real-number arithmetic. He assumes that it is consistent, i.e., that it is free of the possibility of contradictory deductions. Using analytic geometry, he then exhibits a model of EUCLIDEAN GEOMETRY [II.2 §3]. A point is a pair of real numbers; a line is a set of pairs of numbers that satisfy the equation for a line; a circle...; and so on. All of Euclid's axioms are true statements about these

“lines” and “points,” that is, they are true statements about these sets of real numbers. Euclidean geometry is thereby reduced to a fraction of all the true statements about real numbers, and we conclude that if real-number arithmetic is consistent then Euclid’s geometry is consistent. Next Hilbert constructs models of various non-Euclidean geometries in terms of Euclidean geometry, exploring in depth and with great inventiveness which possible axioms follow from which groups of axioms and which are independent yet consistent.

Hilbert was invited to address the Second International Congress of Mathematicians in Paris in 1900. He gave a talk proposing twenty-three problems for the new century. These problems are known today as “Hilbert’s problems”; in a sense they have created a virtual Göttingen where mathematicians have entered into conversation with Hilbert and each other ever since.

Next Hilbert turned to analysis. WEIERSTRASS [VI.44] had found counterexamples to Dirichlet’s principle, which is essentially the assertion that, in variational problems, maxima and minima are always attained. Hilbert proved a modified, but still powerful, version that “salvaged” much of the work that assumed the principle. The larger theme of this period, though, was integral equations and what is now called HILBERT SPACE [III.37]. Newton’s equations for motion are differential equations, and it was natural to phrase equations in physics that way. However, in many cases it was easier to solve problems if the equations were written using integrals rather than derivatives. Between 1902 and 1912 Hilbert attacked a variety of problems from this direction. He viewed the solutions as part of Hilbert space and gave a spectral interpretation analogous to an infinite-dimensional vector space. Thus, an amorphous sea of functions acquired geometric structure.

In 1910 he turned toward mathematical physics and had some successes, but physics was undergoing multiple revolutions and was not ready for mathematical clarification.

When he delivered his problems in 1900, Hilbert was aware that there were contradictions in mathematics as it was then phrased, and specifically in set theory. His second problem asked for a proof that first arithmetic, and then set theory, were consistent. As the debate widened, some mathematicians began to pull back on what they accepted as valid reasoning. Hilbert wanted none of this. By 1918 he was increasingly focused on a program to formally axiomatize mathematics and prove it free of contradictions

using proof-theoretic, combinatorial methods. GÖDEL [VI.92] proved his incompleteness theorems in 1930 and thereby showed that Hilbert’s program, at least as initially conceived, could never be successful. Hilbert was wrong here, but even if wrong, his dream of placing mathematics on a formal foundation stimulated some of the most important work of the twentieth century—and mathematics did not pull back.

Further Reading

Reid, C. 1986. *Hilbert–Courant*. New York: Springer.

Weyl, H. 1944. David Hilbert and his mathematical work. *Bulletin of the American Mathematical Society* 50:612–54.

Benjamin H. Yandell

VI.64 Hermann Minkowski

b. Alexotas, Russia (present day Kaunas, Lithuania), 1864;

d. Göttingen, Germany, 1909

Number theory; geometry; relativity theory

In 1883, the Paris Academy of Sciences awarded its prestigious Grand Prix for mathematical science to the eighteen-year-old student Hermann Minkowski. The prize problem was to give the number of representations of an integer as a sum of five squares of integers. In a manuscript of 140 pages written in German, Minkowski developed a general theory of QUADRATIC FORMS [III.75] that contains the solution to this problem as a special case. Two years later, Minkowski obtained his Ph.D. in Königsberg, and in 1887 he received his habilitation in Bonn with further work on quadratic forms in n variables.

While a student in Königsberg, Minkowski became a close friend of Adolf Hurwitz and HILBERT [VI.63]. In 1894, after Hurwitz had moved to Zürich, Minkowski returned from Bonn to his alma mater, and soon became Hilbert’s successor after Hilbert left for Göttingen. In 1896, Minkowski moved on to Zurich to become Hurwitz’s colleague. In 1902, Hilbert negotiated for another chair of mathematics to be created for Minkowski in Göttingen. There he worked as Hilbert’s colleague and closest friend until he died, unexpectedly, of a ruptured appendix in early 1909.

Minkowski’s later work is characterized by an ingenious use of geometric intuition for the solution of number-theoretic problems. His starting point was a theorem of HERMITE [VI.47] on the smallest positive real that can be represented by a given positive-definite quadratic form of n integer-valued nonzero variables.

By interpreting the quadratic forms in terms of geometric objects such as ellipses (for $n = 2$) or ellipsoids (for $n = 3$), and considering the integer values of the variables as the coordinates of the points of a regular lattice, Minkowski was able to employ the notion of volume to arrive at nontrivial number-theoretic results. His investigations were published in 1896 in a book entitled *The Geometry of Numbers*. Realizing that the geometric arguments based on ellipsoids used only the property of convexity, Minkowski further generalized his theory by introducing a general concept of convex point sets. A *convex body*, according to Minkowski, is one in which the straight line connecting any two interior points lies completely within the set. This notion allowed Minkowski to investigate a geometry in which the Euclidean axiom about the congruence of triangles is replaced by the weaker axiom that the sum of two sides of a triangle is always larger than the third one (which we would nowadays call the *triangle inequality*, the key notion in metric spaces). Theorems about this Minkowskian geometry also produced immediate nontrivial number-theoretic results. Further results were obtained in the theory of CONTINUED FRACTIONS [III.22]. In 1907, Minkowski published introductory lectures on number theory under the title *Diophantine Approximations*.

Minkowski always had a deep interest in physics. In 1906, he wrote the article on capillarity for the authoritative *Encyclopedia of the Mathematical Sciences* (edited by KLEIN [VI.57] and others). In Göttingen, Hilbert and Minkowski gave joint seminars in which they studied contemporary work in electrodynamics by POINCARÉ [VI.61], Einstein, and others. Minkowski soon realized the significance of the fact that the special theory of relativity was a consequence of the invariance of the Maxwell equations under the group of Lorentz transformations (see GENERAL RELATIVITY AND THE EINSTEIN EQUATIONS [IV.13 §1]). He reinterpreted Maxwell-Lorentz electrodynamics geometrically in a mathematical formulation in which no formal distinction between the space and time coordinates exists. This is expressed in the famous opening words of his address to the Cologne meeting of the Society of German Scientists and Physicians a few weeks before his death: "From this hour on, space by itself and time by itself are to sink fully into shadows and only a kind of union of the two should yet preserve autonomy." Minkowski's four-dimensional Lorentz-covariant formulation of special relativity was a prerequisite for Einstein's later general theory of relativity.

Further Reading

- Hilbert, D. 1910. Hermann Minkowski. *Mathematische Annalen* 68:445–71.
 Walter, S. 1999. Minkowski, mathematicians, and the mathematical theory of relativity. In *The Expanding Worlds of General Relativity*, edited by H. Goenner et al., pp. 45–86. Boston: Birkhäuser.

Tilman Sauer

VI.65 Jacques Hadamard

b. Versailles, France, 1865; d. Paris, 1963
Function theory; calculus of variations; number theory; partial differential equations; hydrodynamics

A graduate of the École Normale in Paris, Hadamard obtained a position at the University of Bordeaux in 1893. He returned to Paris in 1897 where he taught at the Collège de France, the École Polytechnique, and the École Centrale until his retirement in 1937. The Hadamard Seminar at the Collège de France, where mathematicians came from around the world to expound on recent results, was an influential and integral part of mathematical life in France between the wars.

Hadamard's first significant papers were concerned with the theory of HOLOMORPHIC FUNCTIONS [I.3 §5.6] of a complex variable, in particular with the analytic continuation of a Taylor series; and in his thesis of 1892 he investigated how the properties of the singularities of a series could be deduced from those of its coefficients. Notably he showed that the radius of convergence R of a Taylor series $\sum a_n z^n$ could be given by $R = (\lim_{n \rightarrow \infty} \sup |a_n|^{1/n})^{-1}$, a result now known as the *Cauchy-Hadamard theorem*. (CAUCHY [VI.29] had published the formula in 1821 but Hadamard, who had discovered it independently, was the first to give a complete proof.) Further results followed, including the famous "Hadamard gap theorem," which gives the condition for the circle of convergence of the series to be a natural boundary of the function. His monograph *La Série de Taylor et son Prolongement Analytique* (1901) proved especially influential. In 1912 he formulated the problem of quasi-analyticity for infinitely differentiable functions.

The year 1892 also saw the appearance of Hadamard's prize-winning memoir on entire functions, in which he used results from his thesis to establish the relations between the coefficients of the Taylor series of an entire function and its zeros, and then applied them to evaluate the genus of the entire function. He applied

this work, and other results from his thesis, to the RIEMANN ZETA FUNCTION [IV.2 §3], which enabled him, in 1896, to prove his most famous result: the PRIME NUMBER THEOREM [V.29]. (The theorem was proved simultaneously by DE LA VALLÉE POUSSIN [VI.67] but in a more complicated way.)

Hadamard's other achievements of the 1890s include a well-known inequality on DETERMINANTS [III.15] (1893), a result essential in the FREDHOLM THEORY [IV.15 §1] of integral equations; and his "three-circles theorem" (1896), which demonstrates the importance of convexity in the study of analytic functions and plays a significant role in interpolation theory.

In 1896 Hadamard won the *Prix Bordin* for his study of the behavior of geodesics on surfaces. (The motivation for studying such geodesics is that they can be used to represent the trajectories of motion in dynamical systems.) It was Hadamard's first major work on a subject other than analysis. His two papers, one on geodesics on a surface of positive curvature (1897) and the other on geodesics on a surface of negative curvature (1898), are characterized by a qualitative analysis inherited from POINCARÉ [VI.61]. The first relies on results from classical differential geometry, while the second is dominated by topological considerations.

Prompted by an interest in the CALCULUS OF VARIATIONS [III.96], Hadamard developed the ideas of Volterra's functional calculus. In 1903 he was the first to describe linear functionals on a function space. By considering the space of continuous functions on a given interval, he showed that every functional is the limit of a sequence of intervals, a result now recognized as a precursor to the RIESZ REPRESENTATION THEOREM [III.18] formulated by RIESZ [VI.74] in 1909. Hadamard's influential *Leçons sur le Calcul de Variations* (1910) is the first book in which the ideas of modern functional analysis can be found.

In applied mathematics Hadamard was primarily concerned with wave propagation, in particular high-speed flows. In 1900 he began working on the theory of partial differential equations, and in 1903 published *Leçons sur la Propagation des Ondes et les Équations de l'Hydrodynamique*; this was followed by *Lectures on Cauchy's Problem in Linear Partial Differential Equations* (1922). The latter contained the details of his fundamental idea of the WELL-POSED PROBLEM [IV.12 §2.4] (i.e., a problem in which the solution must not only exist and be unique but must also depend continuously on the initial data). The origins of the idea can be found in his 1898 paper on GEODESICS [I.3 §6.10].

Hadamard's book *The Psychology of Invention in the Mathematical Field* (1945) is well-known for its discussion of the unconscious and its role in mathematical discovery.

Further Reading

Hadamard, J. 1968. *Collected Works: Œuvres de Jacques Hadamard*, four volumes. Paris: CNRS.

Maz'ya, V., and T. Shaposhnikova. 1998. *Jacques Hadamard. A Universal Mathematician*. Providence, RI: American Mathematical Society/London Mathematical Society.

VI.66 Ivar Fredholm

b. Stockholm, 1866; d. Stockholm, 1927

Professor of Mechanics and Mathematical Physics, Stockholm (1906–27)

In papers of 1900 and 1903 Fredholm solved the integral equations named after him,

$$\varphi(x) + \int_a^b K(x, y)\varphi(y) dy = \psi(x),$$

with a continuous "kernel" K and unknown $\varphi(x)$, by analogy with infinite systems of linear equations and generalized determinants. Both the solution and several ideas attached to it ("Fredholm alternatives") made this work an important stimulus for HILBERT's [VI.63] theory of integral equations (1904–6) and thus a starting point for functional analysis (see OPERATOR ALGEBRAS [IV.15 §1]). The equations arise in the context of problems of mathematical physics, e.g., in potential theory and in the theory of oscillations. Fredholm considered himself primarily a mathematical physicist, and his colleague Mittag-Leffler tried in vain to have him awarded the Nobel Prize for physics.

VI.67 Charles-Jean de la Vallée Poussin

b. Louvain, Belgium, 1866; d. Brussels, 1962

Analytic number theory; analysis

De la Vallée Poussin graduated in engineering (1890) and mathematics (1891) from the Université Catholique de Louvain, where he went on to teach mathematical analysis from 1891 until 1951. His lectures formed the basis for his renowned *Cours d'Analyse Infinitésimale*, which ran to many editions from 1903 to 1959. A member of the most famous academies in Europe and the United States, with honorary doctorates from Paris, Strasbourg, Toronto, and Oslo, he was the first

president (1920) of the International Union of Mathematicians (now the International Mathematical Union). He was made a baron in 1930.

De la Vallée Poussin's main achievement was his proof in 1896 of the PRIME NUMBER THEOREM [V.29] (an asymptotic estimate for the distribution of prime numbers in the integers), first conjectured by GAUSS [VI.26] in around 1793. (The theorem was also proved independently by HADAMARD [VI.65] in the same year, also using complex function theory.) Shortly afterward, de la Vallée Poussin followed his proof with a sharper error term (1899), which he extended to prime numbers in an arithmetic progression.

When LEBESGUE [VI.72] first published his INTEGRAL [III.57] in 1902, de la Vallée Poussin immediately grasped its importance and, using an original approach, described it in the second edition of his *Cours d'Analyse* (1908). In addition, he introduced the concept of the characteristic function of a set (1915), and shortly afterward gave a decomposition theorem for the measure generated by a continuous function of bounded variation (1916).

Of particular importance for approximation theory and the summation of series is de la Vallée Poussin's convolution integral (1908), for approximating periodic functions by trigonometric polynomials. His other significant results in this field include a lower bound for the error in the best approximation of a continuous function by a polynomial (1910), and a convergence test and a summation method for Fourier series (1918).

In 1911 de la Vallée Poussin was responsible for suggesting the Belgian Academy prize question that led to Jackson's and Bernstein's theorems on the order of the best approximation of a continuous function by polynomials. His existence and uniqueness theorem for the Chebyshev problem for an overdetermined system of linear equations (1911) was an important step in LINEAR PROGRAMMING [III.86]; his interpolation formula (1908) was fundamental for sampling theory; and his characterization of new classes of quasi-analytic functions by the rate of decrease of their Fourier coefficients (1915) was a notable development.

De la Vallée Poussin's other achievements include determining a uniqueness condition for multipoint boundary-value problems (1929), which was a significant result for the study of nonoscillatory solutions of linear differential equations; and solving various problems of the conformal representation of multiply connected regions (1930–31). In potential

theory he extended the concept of capacity to arbitrary bounded sets, proved his extraction theorem for bounded sequences of set functions, and, by introducing measure theory into POINCARÉ's [VI.61] "méthode de balayage" ("sweeping-out method") for the DIRICHLET PROBLEM [IV.12 §1], he paved the way for modern abstract potential theory.

Further Reading

Butzer, P., J. Mawhin, and P. Vetro, eds. 2000–4. *Charles-Jean de la Vallée Poussin. Collected Works—Oeuvres Scientifiques*, four volumes. Bruxelles/Palermo: Académie Royale de Belgique/Circolo Matematico di Palermo.

Jean Mawhin

VI.68 Felix Hausdorff

b. Breslau, Germany (now Wrocław, Poland), 1868;

d. Bonn, Germany, 1942

Set theory; topology

Hausdorff studied mathematics at Leipzig, Freiburg, and Berlin between 1887 and 1891, and then started research in applied mathematics at Leipzig under H. Bruns. After his habilitation (1895) he taught first at Leipzig and then later at Bonn (1910–13, 1921–35) and Greifswald (1913–21). He is best known for his work in set theory and general topology, his magnum opus being *Grundzüge der Mengenlehre* ("Basic features of set theory"). It was published in 1914 and had second and third editions in 1927 and 1935. The second edition was so heavily revised in content, however, that it should really be considered a new book.

Hausdorff's early work concentrated on applied mathematics, mainly related to astronomy, in particular the refraction and extinction of light in the atmosphere. He had broad intellectual interests and moved in Nietzschean circles of artists and poets at Leipzig. Under the pseudonym Paul Mongré he wrote two long philosophical essays of which the more prominent was "Das Chaos in kosmischer Auslese" ("The chaos in cosmic selection"). Until 1904 he regularly contributed cultural critical essays to a renowned German intellectual review of the time, continuing to contribute, although less frequently, until 1912. He also published poems and a satirical play.

Hausdorff took up set theory at the turn of the century and gave his first lecture course on the topic in the summer semester of 1901 at Leipzig university. After his turn toward "Cantorism" (set theory) he

began deep and innovative research on order structures and their classification. Among the results of his early work in set theory are the *Hausdorff recursion formula* for exponentiation of cardinals and several contributions to the study of order structures (cofinality, etc.). Although Hausdorff did not pursue active research in the axiomatic foundation of set theory, he contributed important insights on transfinite numbers, in particular a characterization of what are now known as weakly inaccessible cardinals and his *maximal chain principle*, a form of ZORN'S LEMMA [III.1] that predated the latter and differed from it in formulation and intention.

His own contribution to the axiomatic method was oriented toward generalizing classical areas of mathematics and founding them on axiomatic principles within the framework of set theory. Hausdorff's move to use set theory inside mathematics was seminal for the turn toward *modern mathematics* in the sense of the twentieth century, most prominently characterized by the BOURBAKI [VI.96] group. Best known in this respect are his *axiomatization of general topology* in terms of axioms for neighborhood systems, first published in the *Grundzüge* (1914), and the study of the properties of general, or more specialized, TOPOLOGICAL SPACES [III.92]. Less well-known (it remained unpublished until recently) was Hausdorff's *axiomatization of probability theory*, which was presented in a lecture course of 1923 and which preceded KOLMOGOROV's [VI.88] work in this area by about a decade. He also made important contributions to analysis and algebra. In algebra, he contributed to LIE THEORY [III.50] (via what is now called the Baker–Campbell–Hausdorff formula), while in analysis he developed summation methods for divergent series and also a generalization of the Riesz–Fischer theory.

Hausdorff's central goals in using set theory were for applications to analytical disciplines such as function theory. Among his most important contributions in this respect, and of wide-ranging importance, was the concept of HAUSDORFF DIMENSION [III.17], which he introduced to give a notion of dimension to rather general sets (such as, for example, fractal-type sets).

Hausdorff realized that analytical questions of set theory were deeply connected to foundational questions. In 1916 he (and, independently, P. Alexandroff) showed that any uncountable BOREL SET [III.57] in the reals actually has the cardinality of the continuum. This was an important development of a strategy proposed by Cantor to clarify the continuum. Although this strategy did not finally contribute to the decisive results

by Gödel and Cohen on the CONTINUUM HYPOTHESIS [IV.22 §5], it led to the development of an extended field of investigation in the border region between set theory and analysis, now dealt with in DESCRIPTIVE SET THEORY [IV.22 §9]. Hausdorff's second edition of the *Mengenlehre* (1927) was the first monograph in this field.

After the rise to power of the Nazi regime, working conditions and life in general deteriorated more and more drastically for Hausdorff and others of Jewish origin. When Hausdorff, his wife Charlotte, and a sister of hers were ordered to leave their house for local internment in January 1942, they opted for suicide rather than suffering further persecution.

Further Reading

Brieskorn, E. 1996. *Felix Hausdorff zum Gedächtnis. Aspekte seines Werkes*. Braunschweig: Vieweg.

Hausdorff, F. 2001. *Gesammelte Werke einschließlich der unter dem Pseudonym Paul Mongré erschienenen philosophischen und literarischen Schriften*, edited by E. Brieskorn, F. Hirzebruch, W. Purkert, R. Remmert, and E. Scholz. Berlin: Springer.

Hausdorff's voluminous unpublished work (his "Nachlass") is available online at www.aic.uni-wuppertal.de/fb7/hausdorff/findbuch.asp.

Erhard Scholz

VI.69 Élie Joseph Cartan

b. Dolomieu, France, 1869; d. Paris, 1951

Lie algebras; differential geometry; differential equations

Cartan was one of the leading mathematicians of his generation, particularly influential for his work on geometry and the theory of LIE ALGEBRAS [III.50 §§2, 3]. In the bleak years after World War I he was one of the most prominent mathematicians in France. He eventually became a notable influence on the BOURBAKI [VI.96] group, of which his son Henri, another distinguished mathematician, was one of the seven founder members. Cartan held lecturing positions in Montpellier and Lyon before becoming a professor in Nancy in 1903. He went on to gain a lecturing position at the Sorbonne in 1909, becoming a professor in 1912 and remaining there until his retirement.

In his doctoral thesis of 1894 Cartan classified the simple Lie algebras over the field of complex numbers, refining and correcting earlier work of Wilhelm Killing and emphasizing the deep general abstract structures

inherent in the theory. In later years he returned to these ideas and drew out their implications for the study of the corresponding LIE GROUPS [III.50 §1]—these groups have a major bearing on symmetry considerations in physics.

Cartan spent much of his life working on geometry. In the 1870s, and again in the 1890s, KLEIN [VI.57] had analyzed geometry and shown how the major branches (Euclidean, non-Euclidean, projective, and affine) could be unified and treated as special cases of projective geometry. Cartan became interested in the extent to which the group-theoretic ideas that had animated Klein could be adapted to the setting of differential geometry, and especially to spaces of variable CURVATURE [III.80]—the mathematical setting for EINSTEIN'S GENERAL THEORY OF RELATIVITY [IV.13]. In that subject the observations of different observers are related by coordinate transformations, and changes in the gravitational field are expressed through changes in the metric, and hence curvature, of the underlying spacetime manifold. In the 1920s Cartan broadened the setting to what are today called FIBER BUNDLES [IV.6 §5], and showed that Klein's approach could be carried through by concentrating on the possible types of coordinate transformation and the Lie groups to which they can belong.

There are many problems in which one has a multitude of possible observations at each point of a space: for example, the weather at each point of Earth's surface. In Cartan's formulation, Earth's surface is taken as the base MANIFOLD [I.3 §6.9] and the possible observations at each point form another manifold, called the fiber at the point. The pair consisting of all fibers and all points of the base manifold is, roughly, a fiber bundle; the precise concept has proved to be fundamental across the whole field of modern differential geometry. It was to prove a natural setting for the study of what are called *connections* on a manifold, which deal with the way objects, such as vectors, are transformed as they move along curves in the manifold. Cartan's fundamental idea was to capture the symmetry of a geometrical problem by allowing fibers to have a common symmetry group, although aspects of the geometry of the base manifold, such as its curvature, were allowed to vary from point to point in such a way that the base manifold admits no symmetries at all.

Cartan also applied his geometric approach to the study of differential equations, which had earlier been a motivating concern for LIE [VI.53] in the creation of the theory of Lie algebras. He did important work

on systems of equations, and this led him to emphasize the role of what are called exterior forms. Familiar examples include the 1-form (see DIFFERENTIAL FORMS [III.16]) that represents the element of length along a curve, the 2-form that represents the element of area of a surface, and so on. The main thing one does to a 1-form is integrate it; integrating the 1-form that describes arc-length gives length along a curve. Cartan studied systems of equations involving arbitrary 1-forms and was led to discover ways in which the algebra of 1-forms, and more generally the algebra of k -forms for arbitrary k , captures features of the geometry of the manifold on which they are defined. This led him to reformulate a method of studying the geometry of curves and surfaces that had been pursued by Gaston Darboux, the leading French geometer of the previous generation, and to proclaim his method of "moving frames" that again related to the study of fiber bundles and symmetries in differential geometry. This work, together with his work on fiber bundles, remains a major source of ideas for the study of differentiable manifolds to this day.

Further Reading

- Chern, S.-S., and C. Chevalley. 1984. Élie Cartan and his mathematical work. In *Oeuvres Complètes de Élie Cartan*, volume III.2 (1877–1910). Paris: CNRS.
- Hawkins, T. 2000. *Emergence of the Theory of Lie Groups: An Essay in the History of Mathematics, 1869–1926*. New York: Springer.

Jeremy Gray

VI.70 *Emile Borel*

b. Saint-Affrique, France, 1871; d. Paris, 1956

Professor of Mathematics: University of Lille (1893–96), École Normale, Paris (1896–1909); Chair of Theory of Functions (specially created for him), Sorbonne, Paris (1909–41); first director of the Institut Poincaré (1926)

Borel's thesis of 1894 started with problems from within the classical theory of complex functions. With a new theory of MEASURE [III.57] based on CANTOR'S [VI.54] set theory and, in particular, a "covering theorem" (later misnamed the Heine–Borel theorem), he gave a rationale for neglecting certain infinite sets of singularities. He assigned them "measure zero" and thus extended the domain of regularity of the functions considered. Borel's theory of measure, based on operations with infinitely many sets, became widely

known through his influential *Leçons sur la Théorie des Fonctions* (1898) and was later completed and developed into a major tool of analysis by LEBESGUE [VI.72]. It was, in addition, an important prerequisite for the axiomatization of probability by KOLMOGOROV [VI.88].

VI.71 Bertrand Arthur William Russell

b. Trelleck, Wales, 1872; d. Plas Penrhyn, Wales, 1970
Mathematical logic and set theory; philosophy of mathematics

Russell's training at Cambridge University in the early 1890s inspired the part of his long and varied life that relates to mathematics. He divided his Tripos into Part 1 (Mathematics) and Part 2 (Philosophy), and then united these two trainings to seek a general philosophy of mathematics, especially its epistemological foundations, with geometry as the first test case (1897). But over the next few years he changed his philosophical stance, especially when he recognized the significance of CANTOR's [VI.54] set theory from 1896 onward, and also discovered in 1900 a group of mathematicians around PEANO [VI.62] in Turin. Wishing to raise the level of axiomatization and rigor in mathematics, the followers of Peano formalized theories as much as possible, including the "mathematical logic" of propositions and predicates with set theory, but they kept mathematical and logical notions separate. After learning their system and adding to it a logic of relations, Russell decided in 1901 that their distinction of notions was not necessary: *all* notions lay in that logic. This is the philosophical position that has become known as "logicism," and Russell wrote a largely nonsymbolic account of it in *The Principles of Mathematics* (1903). In an appendix to this book he publicized the work of FREGE [VI.56], who had anticipated logicism (but advocated it only for arithmetic and some analysis); Russell read him in detail after forming his own position, which continued to be influenced more by Peano.

Now the job was to expound logicism in Peanesque detail—a daunting task, made even harder by Russell's discovery in 1901 that set theory was susceptible to paradoxes, which would have to be avoided or even solved. He was joined in the effort by his former Cambridge tutor, A. N. Whitehead; eventually three volumes of *Principia Mathematica* appeared between 1910 and 1913. After the basic logic and set theory, the arithmetic of real numbers and also the arithmetic of transfinite numbers were worked out in detail; a fourth volume on geometry was due to be written by Whitehead, but he abandoned it around 1920.

The paradoxes were solved by a "theory of types," which formed a hierarchy of individuals, sets of individuals, sets of sets of individuals, and so on. A set or individual could only be a member of a set immediately above it in the hierarchy; thus, a set could not belong to itself. Comparable restrictions were laid on relations and predicates. While this avoided the paradoxes, it also ruled out a great deal of good mathematics, since different kinds of numbers lay in different types and so could not be brought together for arithmetic operations: for example, $34 + \frac{7}{18}$ was not even definable. The authors proposed the "axiom of reducibility" to allow such definitions to be made; but this was, frankly, just a fudge.

Among the various features of Russell's theory was a form of THE AXIOM OF CHOICE [III.1], called the "multiplicative axiom," that he had found in 1904, just before Ernst Zermelo. It had a curious role within logicism, partly because its logicist status was suspect.

While there was discussion of *Principia Mathematica*, concerning both its logic and its logicism, it tended to be too mathematical for the philosophers and too philosophical for the mathematicians. However, the program influenced some kinds of philosophy, including Russell's own; and as an example of high-level axiomatization it served as a model for foundational studies, including GÖDEL'S INCOMPLETENESS THEOREMS [V.18] of 1931, which showed that logicism as Russell had conceived it could not be achieved.

Further Reading

Grattan-Guinness, I. 2000. *The Search for Mathematical Roots*. Princeton, NJ: Princeton University Press.
 Russell, B. 1983–. *Collected Papers*, thirty volumes. London: Routledge.

Ivor Grattan-Guinness

VI.72 Henri Lebesgue

b. Beauvais, France, 1875; d. Paris, 1941
Theory of the integral; measure; applications in Fourier analysis; dimension in topology; calculus of variations

Lebesgue studied at the École Normale in Paris (1894–97), where he was influenced by the slightly older BOREL [VI.70] and René-Louis Baire. As a teacher at Nancy he completed his seminal thesis "Intégrale, longueur, aire" (1902). After university positions in

Rennes, Poitiers, and at the Sorbonne in Paris, and following war-related research, Lebesgue became a professor at the Sorbonne (1919) and then, finally, at the Collège de France (1921). One year later he was elected to the French Academy of Sciences.

Lebesgue's most important achievement was his generalization of RIEMANN's [VI.49] notion of an integral. This was partly in response to the need to include broader classes of real-valued functions, and partly to give secure foundations to concepts such as the interchangeability of limit and integral in infinite series (particularly Fourier series). Alluding to a famous example (1881) by Vito Volterra of a bounded derivative that could not be integrated, Lebesgue wrote in his thesis:

The kind of integration defined by Riemann does not allow in all cases for the solution of the fundamental problem of the calculus: find a function with a given derivative. It thus seems natural to search for a definition of the integral which makes integration the inverse operation of differentiation in as large a class of functions as possible.

Lebesgue defined his integral by partitioning the range of a function and summing up sets of x -coordinates (or arguments) belonging to given y -coordinates (or ordinates), rather than, as had traditionally been done, partitioning the domain. Lebesgue himself, according to his colleague, Paul Montel, compared his method with paying off a debt:

I have to pay a certain sum, which I have collected in my pocket. I take the bills and coins out of my pocket and give them to the creditor in the order I find them until I have reached the total sum. This is the Riemann integral. But I can proceed differently. After I have taken out all my money I order the bills and coins according to identical values and then I pay the several heaps one after another to the creditor. This is my integral.

The comparison reveals the more theoretical character of Lebesgue's integral, as compared with the more intuitive and natural summation used by Riemann. This meant that more sophisticated functions, which were not necessarily integrable in Riemann's sense, became "summable" according to Lebesgue.

In order to perform his summations, Lebesgue had to base his new integral on Borel's notion of MEASURE [III.57] (1898), which in turn drew heavily on CANTOR's [VI.54] theory of infinite sets. He used infinitely many intervals to cover and to measure sets, and was thus able to measure much less intuitive subsets of the linear continuum (the reals) than had hitherto been con-

sidered. A crucial role was played by the notion of "the set of measure zero" and the consideration of properties that were valid "except for" such sets, i.e., "almost everywhere." This allowed for the theory to be streamlined to include fundamental results such as: "A bounded function is Riemann integrable if and only if the set of its points of discontinuity has measure zero."

Lebesgue completed Borel's theory of measure, making it a true generalization of JORDAN's [VI.52] earlier theory. From Jordan he also borrowed the important notion of a function of bounded variation for his theory of the integral. Lebesgue ascribed a measure to any subset of a "set of measure zero," and opened up broader theoretical questions such as whether there exist any sets that are not Lebesgue-measurable. The latter question was proved in the affirmative by the Italian Giuseppe Vitali in 1905 with the help of the AXIOM OF CHOICE [III.1], while Robert Solovay showed in 1970, with methods of mathematical logic, that without the axiom of choice such existence cannot be proved (see SET THEORY [IV.22 §5.2]). Lebesgue himself remained skeptical about an unlimited use of set-theoretical principles such as the axiom of choice. He held a restrictive view of the "existence" of mathematical objects by making "definability" the touchstone for his empiricist philosophy of mathematics.

Lebesgue's integral—the idea of which was paralleled, although not in such depth, in the work of the English mathematician W. H. Young—served as a sophisticated stimulus to developments in harmonic and functional analysis (e.g., the L^p spaces of RIESZ [VI.74] (1909)). Generalizations to functions defined on n -dimensional space, proposed by Lebesgue himself (1910), contributed to even more general theories of integrals, e.g., the theory of Radon (1913).

Although it took several decades for the importance of Lebesgue's integral to become widely recognized, its significance for applications, especially in the analysis of discontinuous and statistical phenomena of nature and in probability theory, could not be ignored in the long run.

Further Reading

Hawkins, T. 1970. *Lebesgue's Theory of Integration: Its Origins and Development*. Madison, WI: University of Wisconsin Press.

Lebesgue, H. 1972–73. *Œuvres Scientifiques en Cinq Volumes*. Geneva: Université de Genève.

Reinhard Siegmund-Schultze

VI.73 Godfrey Harold Hardy

b. Cranleigh, England, 1877; d. Cambridge, 1947
Number theory; analysis

Hardy was the most influential mathematician in Britain in the twentieth century. With the exception of the years from 1919 to 1931, when he was the Savilian professor of geometry in Oxford, he spent his adult life in Cambridge, where from 1931 until his retirement in 1942 he was the Sadleirian professor of pure mathematics. He became a Fellow of the Royal Society in 1910 and was awarded a Royal Medal in 1920 and the Sylvester Medal in 1940. He died on the day the Royal Society's highest honor, the Copley Medal, was to be presented to him.

At the beginning of the twentieth century, the standard of mathematical analysis was rather low in Britain; Hardy did much to remedy this situation, not only through his research, but also by publishing *A Course of Pure Mathematics* in 1908. This book, which he wrote as "a missionary talking to cannibals," had a tremendous influence on several generations of mathematicians in the United Kingdom. Unfortunately, Hardy's love of pure mathematics, and analysis in particular, somewhat stifled the growth of applied mathematics and algebraic subjects for several decades.

In 1911 he began a long collaboration with LITTLEWOOD [VI.79], with whom he wrote almost one hundred papers: this partnership is generally considered to have been the most fruitful in the history of mathematics. They worked on convergence and summability of series, inequalities, ADDITIVE NUMBER THEORY [V.30] (including Waring's problem and Goldbach's conjecture), and Diophantine approximation.

Hardy was one of the first to do important work on THE RIEMANN HYPOTHESIS [IV.2 §3] when, in 1914, he proved that the zeta function $\zeta(s) = \zeta(\sigma + it)$ has infinitely many zeros on the critical line $\sigma = \frac{1}{2}$ (see LITTLEWOOD [VI.79]). Later, with Littlewood, he proved deep extensions of this result.

From 1914 to 1919 he collaborated with the largely self-taught Indian genius, SRINIVASA RAMANUJAN [VI.82]. They wrote five papers, the most famous of which is about $p(n)$, the number of partitions of n . This is a rapidly growing function: $p(5) = 7$ but

$$p(200) = 3\,972\,999\,029\,388.$$

The GENERATING FUNCTION [IV.18 §§2.4, 3] of $p(n)$, that is,

$$f(z) = 1 + \sum_{n=1}^{\infty} p(n)z^n,$$

is equal to $1/((1-z)(1-z^2)(1-z^3)\cdots)$, so

$$p(n) = \frac{1}{2\pi i} \int_{\Gamma} \frac{f(z)}{z^{n+1}} dz,$$

where Γ is a circle about the origin of radius just less than 1. In 1918, Hardy and Ramanujan not only gave a rapidly convergent asymptotic formula for $p(n)$ but also showed that, for n large enough, $p(n)$ could be calculated *exactly* by taking the integer nearest to the sum of the first few terms. In particular, $p(200)$ can be computed from the first five terms.

Hardy and Ramanujan proved their asymptotic formula for $p(n)$ with the aid of the "circle method"; later, Hardy and Littlewood developed this method into one of the most powerful tools in analytic number theory. In order to estimate contour integrals like the one above, Hardy and Littlewood found it advisable to break up the circle of integration in a subtle way.

Another Hardy-Ramanujan result concerns the number $\omega(n)$ of distinct prime divisors of a "typical" number n . They proved that a "typical" number n has about $\log \log n$ distinct prime factors in a certain precise sense. In 1940 Erdős and Kac sharpened and extended this result by showing that additive number-theoretic functions like $\omega(n)$ obey the GAUSSIAN LAW [III.73 §5] of errors: this gave birth to the important field of probabilistic number theory.

Hardy's name has been attached to several concepts and results, including Hardy spaces, Hardy's inequality, and the HARDY-LITTLEWOOD MAXIMAL THEOREM [IV.11 §3]. For $0 < p \leq \infty$ the *Hardy space* H^p consists of functions analytic in the unit disk that are bounded in various ways; in particular, H^∞ consists of bounded analytic functions. Hardy and Littlewood deduced fundamental properties of H^p from their *maximal theorem*, which relates a function to its "radial limits" at the boundary of the disk. The theory of H^p spaces has found numerous applications not only in analysis, but also in probability theory and control theory.

Hardy and Littlewood loved inequalities of all kinds; their book on the subject with George Pólya, an instant classic the moment it was published in 1934, greatly influenced the development of hard analysis.

Although Hardy was fiercely proud of the purity of his mathematics, in a paper published in 1908 he formulated the extension of the Mendelian law about the

proportions of dominant and recessive characters. This law, which later became known as the *Hardy–Weinberg law*, refuted the idea “that a dominant character should show a tendency to spread over a whole population, or that a recessive should tend to die out.” In a later article he dealt a severe blow to eugenics by giving a simple mathematical argument that showed the futility of forbidding people with “undesirable” characteristics to breed.

In his interest in mathematical philosophy, Hardy was a disciple of RUSSELL [VI.71], whose political views he also shared. He was a secretary of the committee which forced the abolition of the order of merit in the Mathematical Tripos through a reluctant Senate in 1910, and many years later he fought hard for the abolition (not reform!) of the Mathematical Tripos itself, which he considered to be harmful to mathematics in the United Kingdom. After World War I, Hardy led the British efforts to heal the wounds of the international mathematical community, and with the advent of the Nazi persecutions on the Continent in the early 1930s, he was an important figure in an extensive network finding jobs for refugee mathematicians in the United States, Britain, and the Commonwealth. He was a great supporter of the London Mathematical Society: he was not only one of the secretaries for close to twenty years, but also its president for two terms.

Hardy was a militant atheist; as an affectation, he liked to talk of God as his personal enemy. He was a great conversationalist, and was fond of various intellectual games, like putting together cricket teams of bores, bogus poets, Fellows of a Cambridge college, and so on. He loved ball games, especially cricket, baseball, bowls (with the curved woods of his college), and real tennis (as opposed to lawn tennis); to praise people, he frequently likened them to outstanding cricketers.

He had an exceptional gift for collaboration and launching young mathematicians on their research careers. He was a master not only of mathematics, but also of English prose; he was lively and charming, and left a lasting impression even on his casual acquaintances. His poetic book *A Mathematician's Apology*, written toward the end of his life, gives a rare insight into the world of a mathematician.

Further Reading

Hardy, G. H. 1992. *A Mathematician's Apology*, with a foreword by C. P. Snow. Cambridge: Cambridge University Press. (Reprint of the 1967 edition.)

Hardy, G. H., J. E. Littlewood, and G. Pólya. 1988. *Inequalities*. Cambridge: Cambridge University Press. (Reprint of the 1952 edition.)

Béla Bollobás

VI.74 Frigyes (Frédéric) Riesz

b. Győr, Hungary, 1880; d. Budapest, 1956

Functional analysis; set theory; measure theory

After being educated at Budapest University and elsewhere in Europe, Riesz was appointed in 1911 to the University of Kolozsvár (Hungary), which moved in 1920 to become Szeged University; he served twice as Rector. He returned to Budapest in 1946. Most of Riesz's research work lay in mathematical analysis enriched with techniques from set and measure theory, and functional analysis.

One of Riesz's famous results was the converse of a generalization of Parseval's theorem for FOURIER SERIES [III.27]: given a sequence of orthonormal functions on a finite interval, and a sequence a_1, a_2, \dots of real numbers, there exists a function f that can be expanded as a Fourier-type series with respect to those functions with the a_r as coefficients if and only if $\sum_r a_r^2$ is convergent; further, f is itself square summable. He proved the theorem in 1907, simultaneously with the German mathematician Ernst Fischer; so it is named after both of them.

Two years later Riesz found the “representation theorem” named after him. It states that a continuous linear functional that maps continuous functions F over a finite interval I onto the real numbers can be represented as a Stieltjes integral of F over I with respect to a function of bounded variation. It was to be a fertile source of applications and generalizations.

Riesz found these two theorems partly in connection with his study of integral equations, a topic then being developed by HILBERT [VI.63], and partly in connection with his study of functional analysis as formulated by Maurice Fréchet. Hilbert's work had led him to consider infinite matrices, which were then little studied: Riesz wrote the first monograph on them, *Les Systèmes d'Équations Linéaires à une Infinité d'Inconnues* (1913). He also studied the theory of L^p spaces for $p > 1$ (that is, spaces of functions f such that f^p is measure-integrable over some specified interval) and their dual spaces L^q , where $1/p + 1/q = 1$; and he worked on applying his and Fischer's theorem to the self-dual space, now known as HILBERT SPACE [III.37], that is

given by $p = 2$. Later he laid some of the foundations of complete spaces (later known as BANACH SPACES [III.64]), and applied functional analysis to ergodic theory. He summed up much of his work in these areas in the book *Leçons d'Analyse Fonctionnelle* (1952), written with his student B. Szökefalvy-Nagy.

All this work constituted important contributions to theories already laid out in principle by various other mathematicians. Riesz achieved groundbreaking work on subharmonic functions: he modified the DIRICHLET PROBLEM [IV.12§1] by allowing the function that extends a given function into a domain to be subharmonic ("locally less than harmonic") instead of harmonic. He studied some of the applications of these functions to potential theory.

Riesz also studied some foundational aspects of set theory, especially types of ordering, continuity, and generalized Heine–Borel covering theorems. He also reformulated the LEBESGUE INTEGRAL [III.57] in a constructive manner, using step functions and sets of measure zero as primitive notions, and avoiding MEASURE THEORY [III.57] as much as possible.

Further Reading

Riesz, F. 1960. *Oeuvres Complètes*, edited by Á. Császár, two volumes. Budapest: Akadémiai Kiado.

Ivor Grattan-Guinness

VI.75 Luitzen Egbertus Jan Brouwer

b. Overschie, the Netherlands, 1881;
d. Blaricum, the Netherlands, 1966

*Lie groups; topology; geometry; intuitionistic mathematics;
philosophy of mathematics*

Brouwer entered the University of Amsterdam at the age of sixteen, where his teacher was D. J. Korteweg. The young Brouwer taught himself modern mathematics, as well as a fair amount of philosophy. As a graduate student he published some original papers on the decomposition of rotations in four-dimensional space. He also published a brief monograph on mysticism that contained a number of ideas that became prominent in his later philosophy. In his dissertation of 1907 he solved a special case of HILBERT's [VI.63] fifth problem (the elimination of differentiability conditions from the axioms of LIE GROUPS [III.50§1]), and he presented his first program for "constructive mathematics."

The basis of his mathematics was the *ur-intuition of mathematics*: the continuum and the natural numbers are simultaneously created from intuition. Mathematical objects (including proofs) are mental creations. After sketching the development of the basic parts of mathematics, Brouwer went on to criticize contemporary mathematics for transcending the bounds of the human mind. In particular, he criticized CANTOR [VI.54] for introducing sets beyond human recognition, and Hilbert for the axiomatic method and for formalism. He criticized the latter's consistency program and denied that "consistency implies existence."

In his 1908 paper "The unreliability of the logical principles," Brouwer explicitly rejected the principle of the excluded middle as unreliable (and also rejected Hilbert's dogma that "all mathematical problems can be solved in one way or another"). Between 1909 and 1913 Brouwer worked in topology. He continued his work on Lie groups, and noted that topology (in the style of Cantor–Schoenflies) was in need of a sound basis. In his paper "Zur Analysis Situs" (1910), he spelled out a number of notions and examples (curves, indecomposable continua, three domains with one common boundary). This was the beginning of his revision of set-theoretic topology. At the same time he started two lines of research: one on homeomorphisms from surfaces to themselves, establishing FIXED-POINT THEOREMS [V.13] on the sphere and the *plane translation theorem* (a characterization of fixed-point free homeomorphisms of the Euclidean plane); and one on vector distributions on the sphere, yielding existence theorems for singular points, and a characterization of these points. The best-known theorem in this area is Brouwer's "hairy ball theorem" (no matter how one combs a hairy ball, there is always a crown). In 1910 Brouwer published a direct topological proof of the Jordan curve theorem, which remains one of the most elegant proofs. The so-called new topology opened with Brouwer's "invariance of dimension" theorem (1910). He then laid the basis for topology of MANIFOLDS [I.3§6.9], where his basic tool was the Brouwer degree of continuous mappings. The basic paper is his "Über Abbildungen von Mannigfaltigkeiten" ("On mappings of manifolds," 1911), which contained most of the tools for the new topology, e.g., simplicial approximation, mapping degree, HOMOLOGY [IV.6§§2, 3], singularity index (in his own terminology), and also the fundamental properties of the new notions.

Brouwer's new topological insights and techniques led him to a wealth of spectacular results: the Brouwer

fixed-point theorem, the invariance-of-domain theorem, the higher-dimensional Jordan theorem, and the definition of dimension, including the soundness proof (that \mathbb{R}^n has dimension n). He also applied his invariance-of-domain theorem to the theory of automorphic functions and uniformization, thus proving the correctness of the *Klein-Poincaré continuity method* (1912).

During World War I Brouwer returned to the foundations of mathematics; he conceived his mature INTUITIONISTIC MATHEMATICS [II.7 §3.1], which fully exploited the potential of constructive mathematics, based on mentally created objects and notions. The key notions were (infinite) choice sequences (i.e., sequences determined by more or less free choices (by the mathematician) of mathematical objects, say natural numbers), well-orderings, and intuitionistic logic. In “Brouwer’s universe” strong results can be obtained: the “continuity principle,” for example, which says that a function that assigns natural numbers to choice sequences is continuous (i.e., the output is determined by a finite piece of the (infinite) input); and certain transfinite induction principles, in particular the novel principle of “bar-induction.” With the help of these principles he showed that (i) all real functions on a closed segment are uniformly continuous and (ii) the continuum is indecomposable (cannot be split). This enabled him to refute the principle of the excluded middle in a strong sense: it is not the case that each real number is zero or nonzero. In Brouwer’s universe many classical theorems, such as the intermediate-value theorem and the Bolzano-Weierstrass theorem, fail.

Brouwer’s mathematical universe lacked the logical “principle of the excluded middle,” but instead it had certain strong constructive principles at its disposal, which turned it into an alternative to the traditional universe, with a comparable strength.

His foundational program brought him into conflict with Hilbert (“intuitionism versus formalism”). In 1928 matters came to a head and, in an incident famously described by Einstein as “the war of frogs and mice,” Hilbert succeeded in getting Brouwer removed (after fourteen years’ service) from the editorial board of *Mathematische Annalen*.

Brouwer was unconventional and had wide-ranging interests: art, literature, politics, philosophy, mysticism. He was a staunch internationalist.

He was a professor at the University of Amsterdam from 1912 until 1951.

Further Reading

- Brouwer, L.E.J. 1975–76. *Collected Works*, two volumes. Amsterdam: North-Holland.
 van Dalen, D. 1999–2005. *Mystic, Geometer and Intuitionist. The Life of L.E.J. Brouwer*, two volumes. Oxford: Oxford University Press.

Dirk van Dalen

VI.76 Emmy Noether

b. Erlangen, Germany, 1882; d. Bryn Mawr, Pennsylvania, 1935
Algebra; mathematical physics; topology

Noether began her career with a feat of classical algebra, which she transmuted into the NOETHER CONSERVATION THEOREMS [IV.12 §4.1] for physics. She became a founder of modern abstract algebra and the leader in spreading that algebra all across mathematics.

Her father Max Noether and family friend Paul Gordan were Erlangen mathematicians and favored educating women. Gordan made heroic calculations of invariants in algebra. A quadratic polynomial $Ax^2 + Bx + C$ has essentially just one invariant, the discriminant $\sqrt{B^2 - 4AC}$ used in the quadratic formula. As Gordan’s student, Noether found 331 independent invariants of degree-four polynomials in three variables, and proved that all others depend on them. It was impressive, though not, as it turned out, groundbreaking.

HILBERT [VI.63] brought her to Göttingen in 1915 to work on invariants for differential equations in general relativity by reducing them to algebra. That year she found her conservation theorems, which show that the conserved quantities of a physical system correspond to its symmetries. For example, if a system has laws unchanging with time, so that a time shift is a symmetry of the system, then energy is conserved in the system (Feynman 1965, chapter 4). These theorems became fundamental in Newtonian physics and especially quantum mechanics. They also showed that general relativity admits conservation laws only in special cases.

Noether saw the creation of general abstract algebra as her life’s work. Instead of classical algebra with real numbers, or complex numbers, and polynomials using them, she would study any system satisfying abstract rules such as the RING AXIOMS [III.83] or the GROUP AXIOMS [I.3 §2.1]. Concrete examples include the ring of all algebraic functions defined on a space (such as a sphere), and the group of all symmetries

of a given space. She largely created the now-standard style of abstract algebra. Her ideas were also adopted in ALGEBRAIC GEOMETRY [IV.4] where every abstract ring appears as the ring of functions on a corresponding space called a SCHEME [IV.5 §3].

She turned her attention away from operations on elements of a system, like plus and times, and focused instead on relating whole systems to each other, such as rings R, R' related by ring HOMOMORPHISMS [I.3 §4.1] from R to R' . She organized all algebra around her homomorphism and isomorphism theorems. Her aim was to show how IDEALS [III.83 §2] and their corresponding homomorphisms could replace equations between elements as the basic tools for stating and proving theorems. (This approach was to come to fruition in the 1950s, with the advent of Grothendieck-style homological algebra.)

Topologists studied TOPOLOGICAL SPACES [III.92] by looking at continuous functions from one space to another. Noether saw how her algebraic methods could apply here, and convinced young topologists in the 1920s to use them in algebraic topology. Each topological space S has HOMOLOGY GROUPS [IV.6 §4] $H_n S$ with the property that continuous functions from S to S' induce group homomorphisms from $H_n S$ to $H_n S'$. Theorems of topology follow by abstract algebra. This relationship between homomorphisms and continuous functions is what inspired CATEGORY THEORY [III.8].

In the 1930s Noether pursued the algebra of GALOIS THEORY [V.24] through a radically simplified abstract theory of groups acting on rings. The applications are quite arcane, beginning with CLASS FIELD THEORY [V.31] and eventually growing into group cohomology and many other algebraic and topological methods used in ARITHMETIC GEOMETRY [IV.5].

Exiled from Germany by the Nazis in 1933, she died following surgery in the United States, at the height of her creative power.

Further Reading

- Brewer, J., and M. Smith, eds. 1981. *Emmy Noether: A Tribute to Her Life and Work*. New York: Marcel Dekker.
- Feynman, R. 1965. *The Character of Physical Law*. Cambridge, MA: MIT Press.

Colin McLarty

VI.77 Wacław Sierpiński

b. Warsaw, 1882; d. Warsaw, 1969

Number theory; set theory; real functions; topology

Sierpiński studied mathematics at the Russian university in Warsaw under the guidance of Georgii Voronoi. In his first paper (1906), he improved GAUSS's [VI.26] estimate for the difference between the number of lattice points inside the circle $x^2 + y^2 \leq N$ and the area of the circle, showing that it is $O(N^{1/3})$.

He became an associate professor at the University of Lwów in 1910, at which point his interest shifted to set theory, on which he wrote a textbook in 1912, only the fifth book ever to be published on the subject. His first important results on set theory were obtained during World War I, which he spent in Russia: in 1915–16 he constructed two curves that were among the first published examples of fractals, one known now as Sierpiński's gasket and the other as Sierpiński's carpet. The latter is the set of all points (x, y) in the square $[0, 1]^2$ such that, when written out as base 3 decimals, there is no position in which both x and y have a 1. It is also known as Sierpiński's universal curve, since it contains a homeomorphic image of every planar continuum (a continuum is a compact connected set) without interior points.

In 1917, Souslin had shown that projections of BOREL SETS [III.57] (from the plane into the line, say) need not be Borel. Together with Lusin, Sierpiński proved in 1918 that in fact every analytic set (a projection of a Borel set) is the intersection of \aleph_1 Borel sets (where \aleph_1 is the smallest uncountable cardinal). That same year he also published an important study of THE AXIOM OF CHOICE [III.1] and the role it plays in set theory and analysis, and proved that no continuum can be decomposed into countably many pairwise disjoint nonempty closed subsets.

In 1919 Sierpiński was made a full professor at the new Polish University of Warsaw and in 1920 he founded (together with Janiszewski and Mazurkiewicz) the first specialized mathematical journal, *Fundamenta Mathematicae*, which was devoted to set theory, topology, and applications. He remained its editor until 1951. Among his results published in volume 1 are a proof that every countable subset of \mathbb{R}^n without isolated points is homeomorphic to the rationals; a complete classification of countable compact subsets of \mathbb{R}^n , obtained jointly with Mazurkiewicz; and a neces-

sary and sufficient condition for a subset of \mathbb{R}^n to be a continuous image of an interval.

Using the CONTINUUM HYPOTHESIS [IV.22 §5] ($\aleph_1 = 2^{\aleph_0}$), he constructed an UNCOUNTABLE SET [III.11] of reals, now known as a *Sierpiński set*, such that every uncountable subset of it is nonmeasurable (1924); and also a one-to-one mapping of the line into itself that maps sets of MEASURE ZERO [III.57] to sets of first category, in such a way that every set of the first category is obtained (1934). The former result is highly paradoxical (no explicit example of a nonmeasurable set is known); the latter has led, thanks to Erdős, to the following duality principle. Let P be any proposition involving solely the notions of measure zero, first category, and pure set theory. Let P^* be the proposition obtained from P by interchanging the terms “set of measure zero” and “set of first category.” Then P and P^* are equivalent, assuming the continuum hypothesis.

Sierpiński wrote a monograph devoted to the continuum hypothesis in 1934, entitled *Hypothèse du Continu*. Together with TARSKI [VI.87] he introduced the notion of STRONGLY INACCESSIBLE CARDINALS [IV.22 §6] (1930), meaning cardinals m that cannot be obtained as products of fewer than m cardinals less than m . He also worked in Ramsey theory, giving a limitation on infinite extensions to Ramsey’s theorem. To be precise, Ramsey had proved that, whenever one finitely colors the pairs from the natural numbers, there is an infinite monochromatic subset (i.e., a subset all of whose pairs have the same color); Sierpiński showed that, by contrast, one can 2-color the pairs from a ground set of size \aleph_1 in such a way that there is no monochromatic subset of size \aleph_1 . He also deduced the axiom of choice from the generalized continuum hypothesis (formulated without cardinals in 1947).

In his old age he returned to number theory and became the editor of *Acta Arithmetica* (1958–69).

Further Reading

Sierpiński, W. 1974–76. *Oeuvres Choiesies*. Warsaw: Polish Scientific.

Andrzej Schinzel

VI.78 George Birkhoff

b. Oversiel, Michigan, 1884; d. Cambridge, Massachusetts, 1944
Difference equations; differential equations; dynamical systems; ergodic theory; relativity theory

At the International Congress of Mathematicians in 1924 the Russian mathematician A. N. Krylov described Birkhoff as “the POINCARÉ [VI.61] of America.” It was an apt description and one that Birkhoff would have relished, for he was deeply influenced by Poincaré’s work, in particular his great treatise on celestial mechanics.

Birkhoff studied first at Chicago under E. H. Moore and Oskar Bolza, and then at Harvard under W. F. Osgood and Maxime Bôcher. Returning to Chicago, he was awarded his doctorate in 1907 for a thesis on asymptotic expansions, boundary-value problems, and Sturm-Liouville theory. In 1909, after two years at Wisconsin under E. B. Van Vleck, he went to Princeton, where he formed a close association with Oswald Veblen. In 1912 he moved to Harvard and remained there, in professorial positions, until his sudden death in 1944. Birkhoff was steadfast in his support for the development of American mathematics, supervising forty-five doctoral students, including Marston Morse and Marshall Stone, and holding many distinguished positions within the scientific community. He was generally recognized, both at home and abroad, as the leading American mathematician of his generation.

Birkhoff first came to prominence with a memoir on the theory of linear difference equations (1911), and he continued to publish on the topic intermittently throughout his career. Related to this work were several papers on the theory of linear differential equations and a paper on the generalized Riemann problem (1913), which concerns complex functions defined by differential equations. (Until recently it was believed that the latter paper included a solution to Hilbert’s twenty-first problem, the Hilbert-Riemann problem, but in 1989 Bolibruh proved this belief to be mistaken.)

Throughout his life Birkhoff’s deepest interest in analysis lay in DYNAMICAL SYSTEMS [IV.14] and it was here that he enjoyed his greatest success. His overarching aim was to obtain a reduction of the most general dynamical system to a normal form from which a complete qualitative characterization could be deduced. As with Poincaré, the study of periodic motions was central to his work, and he wrote extensively on the THREE-BODY PROBLEM [V.36] as well as on questions connected with stability. Of his memoir on dynamical systems with two degrees of freedom (1917), which won the Bôcher prize in 1923, he is said to have remarked that it was as good a piece of work as he was ever likely to do. Another celebrated achievement was

his proof of Poincaré's topological "last geometric theorem," the publication of which brought him immediate international acclaim (1913). (The theorem states that any one-to-one area-preserving transformation of an annulus that moves the boundary circles in opposite directions must have at least two fixed points, and its importance lies in the fact that its proof implies the existence of periodic orbits in the restricted three-body problem.) He introduced several new concepts into dynamical theory, including "recurrent motion" (1912) and "metric transitivity" (1928), and promoted the use of symbolism in dynamics (1935), the latter helping to pave the way for the formalized development of symbolic dynamics (the branch of dynamical systems invented by HADAMARD [VI.65] (1898) that deals with spaces consisting of infinite sequences of symbols) by Marston Morse and Gustav Hedlund at the end of the 1930s. His book *Dynamical Systems* (1927) was the first book on the qualitative theory of systems defined by differential equations. Awash with topological ideas, it provides a connected account of much of his earlier research.

Closely related to Birkhoff's dynamical research was his work on ERGODIC THEORY [V.11]. Stimulated by the theorems of Bernard Koopman and VON NEUMANN [VI.91], Birkhoff presented his own ergodic theorem in 1931, a fundamental result both for statistical mechanics and for MEASURE THEORY [III.57], the proof of which combined Poincaré's topological approach with the use of Lebesgue measure theory. (Roughly speaking, Birkhoff's ergodic theorem states that for any dynamical system given by differential equations that possesses an invariant volume integral, there is a definite "time probability" p that any moving point, except those of a set of measure zero, will be in an assigned region v . In other words, if t is a total elapsed time interval and t^* is the portion of time during which the point is in v , then $\lim t^*/t = p$.)

In the creation of physical theories Birkhoff advocated mathematical symmetry and simplicity above physical intuition. His books on relativity theory (which were among the first on the subject in English), *Relativity and Modern Physics* (1923) and *The Origin, Nature, and Influence of Relativity* (1925), were characteristically original and widely read. At the time of his death he was engaged in developing a new theory of matter (taken to be a perfect fluid), electricity, and gravitation, which he had first proposed in 1943 and which, unlike Einstein's theory, was based on flat spacetime.

Birkhoff published in several other fields, including the CALCULUS OF VARIATIONS [III.96] and map coloring, and he was the coauthor (with Ralph Beatley) of a textbook of elementary geometry (1929). His paper (with O. D. Kellogg) on fixed points in function space (1922) provided a stimulus for the later work of Leray and Schauder.

Birkhoff had a lifelong interest in the arts and was fascinated by the problem of analyzing the fundamentals of musical and artistic form. In later life he lectured extensively on the application of mathematics to aesthetics, and his book *Aesthetic Measure* (1933) enjoyed popular success.

Further Reading

Aubin, D. 2005. George David Birkhoff. Dynamical systems. In *Landmark Writings in Western Mathematics 1640–1940*, edited by I. Grattan-Guinness, pp. 871–81. Amsterdam: Elsevier.

VI.79 John Edensor Littlewood

b. Rochester, England, 1885; d. Cambridge, England, 1977
Analysis; number theory; differential equations

Littlewood made important contributions to many branches of analysis and analytic number theory, including Abelian and Tauberian theory, the RIEMANN ZETA FUNCTION [IV.2 §3], WARING'S PROBLEM, GOLDBACH'S CONJECTURE [V.30], harmonic analysis, probabilistic analysis, and nonlinear differential equations. He loved concrete problems such as THE RIEMANN HYPOTHESIS [IV.2 §3]: he was arguably the best problem solver of his generation. Much of his work was done in collaboration with HARDY [VI.73]: the Hardy-Littlewood partnership dominated the mathematical scene in the United Kingdom for a third of a century. With the exception of three years in Manchester, he lived all his adult life in Trinity College, Cambridge. From 1928 until his retirement in 1950, he was the first holder of the Rouse Ball Chair of Mathematics in Cambridge.

His first major result, published in 1911, was a deep converse of ABEL's [VI.33] classical theorem that if a series of reals $\sum a_n$ sums to A , then $\sum a_n x^n$ also tends to A as $x \rightarrow 1$ from below. In general, the converse is false, but Tauber had proved that it is true if $na_n \rightarrow 0$. Littlewood extended this by weakening the condition to na_n being bounded. This result gave rise to an extended area of analysis called Tauberian theorems.

In the theory of functions, he did elegant, important, and innovative work on injective holomorphic functions, the minimum modulus, and subharmonic functions. In particular, he worked on the conjecture that Bieberbach made in 1916 that if $f(z) = z + a_2 z^2 + a_3 z^3 + \cdots$ is an injective HOLOMORPHIC FUNCTION [I.3 §5.6] in the open disk $\Delta = \{z : |z| < 1\}$, then $|a_n| \leq n$ for every n . Littlewood proved in 1923 that $|a_n| < en$ for every n . After many improvements by a number of people, the constant e was eventually reduced to a value close to 1, before de Branges proved the full conjecture in 1984.

Littlewood had a lifelong interest in the zeta function. This is defined in the half-plane $\operatorname{Re}(s) > 1$ by the absolutely convergent series

$$\zeta(s) = \zeta(\sigma + it) = \frac{1}{1^s} + \frac{1}{2^s} + \frac{1}{3^s} + \cdots,$$

and in the whole complex plane by analytic continuation. In fact, the second problem suggested to him by his supervisor was the Riemann hypothesis that the zeros of $\zeta(s)$ in the “critical strip” $0 < \sigma < 1$ are on the “critical line” $\sigma = \frac{1}{2}$. If true, this famous conjecture would imply deep results about the distribution of primes. Most of Littlewood’s work on the zeta function was done in collaboration with Hardy and concerned analytic properties of $\zeta(s)$.

In addition to his work with Hardy, he also made use of the zeta function to prove a striking theorem about the error term in the PRIME NUMBER THEOREM [V.29]. The prime number theorem itself had been proved by HADAMARD [VI.65] and, independently, by DE LA VALLÉE POUSSIN [VI.67] in 1896. This fundamental result states that $\pi(x)$, the number of primes less than x , is asymptotic to the “logarithmic integral” $\operatorname{li}(x) = \int_0^x (1/\log t) dt$. There was much numerical evidence that $\pi(x) < \operatorname{li}(x)$ for all x ; in particular, by 1914 this inequality was known to hold for all $2 \leq x \leq 10^7$. Nevertheless, Littlewood proved that $\operatorname{li}(x) - \pi(x)$ changes sign infinitely often. Interestingly, he did not obtain any explicit bound for a value x with $\pi(x) > \operatorname{li}(x)$; the first such bound, given by Skewes in 1955, was

$$10^{10^{10^{1000}}}.$$

Hardy and Littlewood proved important approximate formulas for $\zeta(s)$, which they used to deduce that, in a certain sense, $\zeta(s)$ is “small” on the critical line; this was viewed as a breakthrough. Littlewood also studied the number of zeros of $\zeta(s)$ in a rectangle $0 < \sigma < 1$, $0 < t \leq T$.

In 1770, in his *Meditationes Algebraicae*, WARING [VI.21] asserted on the basis of empirical evidence that every natural number is the sum of nine nonnegative integral cubes, nineteen fourth powers, and so on: for every natural number k there is a minimal integer $g(k)$ such that every natural number is the sum of $g(k)$ nonnegative k th powers. In 1909 HILBERT [VI.63] used complicated algebraic identities to prove that $g(k)$ indeed exists, but the bounds he obtained on $g(k)$ were rather weak. In the 1920s, in a groundbreaking series of papers entitled *Partitio Numerorum*, Hardy and Littlewood introduced an analytic method that could be used to tackle not only Waring’s problem of determining $g(k)$, but many other problems as well. The origins of this “circle method” of Hardy and Littlewood go back to the work of Hardy and RAMANUJAN [VI.82] on the partition function, but the technical difficulties that Hardy and Littlewood had to overcome were much greater than in that earlier work. This method enabled them to show, for example, that every sufficiently large number is the sum of nineteen fourth powers. (In 1986, Balasubramanian, Dress, and Deshouillers proved that $g(4)$ is indeed 19.) More importantly, they gave an asymptotic estimate for the number of representations of n as a sum of at most s positive k th powers.

The circle method also provides a possible line of attack on *Goldbach’s conjecture* that every even number greater than two is the sum of two primes, and gives strong heuristic evidence for the strengthened version of the *twin-prime conjecture* that the number of primes $p \leq n$ such that $p + 2$ is also a prime is asymptotic to $c \int_2^n (1/(\log t)^2) dt$ for a constant $c > 0$. The so-called *k-tuple conjecture* of Hardy and Littlewood is a further extension of this conjecture for “constellations of primes.”

Much of Littlewood’s remarkable work on harmonic analysis was done in collaboration with R.E.A.C. Paley in the early 1930s. The starting point of the LITTLEWOOD–PALEY THEORY [VII.3 §7] is an inequality concerning trigonometric polynomials. Roughly speaking, Littlewood and Paley related the size of a function to the projection of its FOURIER COEFFICIENTS [III.27] onto various intervals. The original one-dimensional Littlewood–Paley theory has been extended to higher dimensions, arbitrary intervals, and even to tensors on two-dimensional compact manifolds; the theory has connections to such varied topics as WAVELETS [VII.3], semigroups acting on L^p -spaces of functions with values in a BANACH SPACE [III.64], and the geometry of null hypersurfaces for rough Einstein metrics.

Littlewood was also a formidable applied mathematician. During World War I he worked on ballistics, and during World War II, with his collaborator Mary Cartwright, he worked on the van der Pol oscillator in order to help the development of radio. Cartwright and Littlewood were among the first to combine topological and analytical methods to tackle differential equations, and discovered many of the phenomena that later became known as “chaos”: they proved that chaos could arise even in equations originating in real engineering problems.

From 1910 until his death sixty-seven years later, Littlewood lived in the same set of spacious rooms in Trinity College, Cambridge. He was a great raconteur: after almost every dinner he was to be found in the Combination Room drinking claret in the company of Fellows and any mathematicians who might be visiting. In spite of his tremendous output, he suffered for decades from severe bouts of depression, from which he was cured only in 1957. He practiced his belief that mathematicians should take a vacation of at least twenty-one days a year during which they should do no mathematics. He was a keen and skilled rock climber and an avid Alpine skier. Although not an active musician, on most days he listened to Bach, Beethoven, and Mozart for hours on end.

In 1943, when he was awarded the Sylvester Medal of the Royal Society, the citation read: “Littlewood, on Hardy’s own estimate, is the finest mathematician he has ever known. He was the man most likely to storm and smash a really deep and formidable problem; there was no one else who could command such a combination of insight, technique and power.”

Further Reading

Littlewood, J. E. 1986. *Littlewood’s Miscellany*, edited and with a foreword by B. Bollobás. Cambridge: Cambridge University Press.

Béla Bollobás

VI.80 Hermann Weyl

b. Elmshorn, Germany, 1885; d. Zürich, 1955
Analysis; geometry; topology; foundations; mathematical physics

Weyl studied mathematics at Göttingen under HILBERT [VI.63], KLEIN [VI.57], and MINKOWSKI [VI.64] between 1904 and 1908. His first teaching positions were at Göttingen (1910–13) and ETH Zürich (1913–30). In 1930 he

accepted the call to Göttingen as Hilbert’s successor. After the rise to power of the Nazis, he emigrated to the United States and became a member of the newly founded Institute of Advanced Studies at Princeton (1933–51).

Weyl made contributions to real and complex analysis, geometry and topology, LIE GROUPS [III.50 §1], number theory, the foundations of mathematics, mathematical physics, and philosophy. He contributed at least one book to each of these fields, publishing thirteen in total. Together with his other technical and conceptual innovations, these books were all of lasting influence: many had a pronounced and immediate effect.

His early research dealt with integral operators and differential equations with singular boundary conditions. His fame came later, with his book *The Concept of a Riemann Surface* (1913). This grew out of a lecture course in the winter of 1910–11 and built upon Klein’s intuitive treatment of RIEMANN’S [VI.49] geometric function theory and Hilbert’s justification of the DIRICHLET PRINCIPLE [IV.12 §3.5]. Here Weyl gave a new presentation of the properties of RIEMANN SURFACES [III.81], which became highly influential for the geometric function theory of the twentieth century.

His second book, *The Continuum* (1918), marked the beginning of Weyl’s interest in the foundations of mathematics. He was critical of Hilbert’s “formalist” program for an axiomatic foundation of mathematics, and explored the possibility of a semi-formalized arithmetical approach to a strictly constructivist foundation of real analysis. Shortly thereafter he shifted toward BROUWER’S [VI.75] intuitionistic program and attacked Hilbert’s foundational views even more strongly in a famous article of 1921. In the late 1920s he developed a more balanced view of the foundational questions. After World War II he returned to a weak preference for his arithmetical constructive approach of 1918.

At the same time as he was working on foundational questions Weyl took up Einstein’s theory of general relativity and wrote his third book, *Space–Time–Matter*. It was first published in 1918, and appeared in five successive editions until 1923. This was one of the first monographs on relativity theory and was among the most influential. The book represented only the tip of the iceberg of his contributions to differential geometry and general relativity. Weyl undertook this research within a broad conceptual and philosophical framework. One of the outcomes of this approach was his *Analysis of the Problem of Space* (1923), in which he sketched ideas that would later be analyzed in terms of

the geometry of FIBER BUNDLES [IV.6 §5] and the study of *gauge fields*. He had already introduced gauge fields (and the key idea of a point-dependent rescaling of the metric) in 1918 for a generalization of RIEMANNIAN GEOMETRY [I.3 §6.10] and a geometrically unified field theory of gravity and electromagnetism.

Weyl made his most influential contributions to pure mathematics around the middle of the 1920s with his work on the REPRESENTATION THEORY [IV.9] of semisimple Lie groups. Combining CARTAN's [VI.69] insights into the representations of LIE ALGEBRAS [III.50 §2] with methods developed by Hurwitz and Schur, Weyl used his knowledge of the topology of manifolds and developed the core of the general theory of representations of Lie groups in a blend of geometric, algebraic, and analytic methods. He extended and refined this work and it formed the core of his later book *The Classical Groups* (1939)—a harvest of his work and lectures on this topic during his Princeton years.

Along with all this work, Weyl actively followed the rise of the new quantum mechanics. In 1927–28 he gave a lecture course at ETH on the topic, which gave rise to his next book on mathematical physics, *Group Theory and Quantum Mechanics* (1928). Weyl emphasized the conceptual role of group methods in the symbolic representation of quantum structures, in particular the intriguing interplay between representations of the special linear group and PERMUTATION GROUPS [III.70]. A second step in his gauge theory of the electromagnetic field was published separately, which gave rise to a modified gauge theory of electromagnetism. This was endorsed by leading theoretical physicists, including Pauli, Schrödinger, and Fock. It served as a starting point for the next generation of physicists who developed gauge field theories in the 1950s and 1960s.

Weyl's research in mathematics and physics was shaped by his philosophical outlook and he included his philosophical reflections on scientific activity in many of his publications. Most influential was his contribution to a philosophical handbook, *Philosophy of Mathematics and Natural Science*, originally published in German in 1927 and translated into English in 1949. It became a classic in the philosophy of science.

Further Reading

Chandrasekharan, K., ed. 1986. *Hermann Weyl: 1885–1985. Centenary Lectures delivered by C. N. Yang, R. Penrose, and A. Borel at the Eidgenössische Technische Hochschule Zürich*. Berlin: Springer.

Deppert, W., K. Hübner, A. Oberschelp, and V. Weidemann, eds. 1988. *Exact Sciences and Their Philosophical Foundations*. Frankfurt: Peter Lang.

Hawkins, T. 2000. *Emergence of the Theory of Lie Groups. An Essay in the History of Mathematics 1869–1926*. Berlin: Springer.

Scholz, E., ed. 2001. *Hermann Weyl's Raum-Zeit-Materie and a General Introduction to His Scientific Work*. Basel: Birkhäuser.

Weyl, H. 1968. *Gesammelte Abhandlungen*, edited by K. Chandrasekharan, four volumes. Berlin: Springer.

Erhard Scholz

VI.81 Thoralf Skolem

b. Sandsvaer, Norway, 1887; d. Oslo, 1963
Mathematical logic

Thoralf Skolem was one of the major logicians of the twentieth century, often a lone voice in his understanding of the subtle relationship between abstract set theory and logic. He also worked on Diophantine equations and on group theory, but his contributions to mathematical logic have proved the most lasting. He taught at Bergen and Oslo, was for a time President of the Norwegian Mathematical Society and an editor of its journal, and in 1954 was named a Knight of the First Class in the Royal Order of St. Olav by the king of Norway.

In 1915 Skolem extended a result obtained by the Polish mathematician Leopold Löwenheim. His conclusion (published in 1920 and known as the Löwenheim–Skolem theorem) says that if a mathematical theory defined using only the first-order predicate calculus has a model, then it has a countable model. Here a model is a set of mathematical objects that obeys the axioms of the theory (see LOGIC AND MODEL THEORY [IV.23]). Now, the real numbers are definable in such a theory (for example, ZERMELO–FRAENKEL SET THEORY [IV.22 §3], or any other axioms for set theory). From this we obtain the so-called Skolem paradox, that the real numbers can be defined in a theory with a countable model, even though it had been known since the time of CANTOR [VI.54] that the real numbers are uncountable. How can this paradox be resolved?

The answer is that one has to be very careful about what we mean by “countable.” In this strange countable model of set theory, we can see that the reals are countable, but *to the model* the reals may be uncountable. In other words, the actual enumeration of the reals

that we can see (i.e., the actual bijection between the reals and the natural numbers) may not belong to the model: the model can be so “small” that it is missing some functions. Skolem’s paradox highlights the difference between the viewpoint from *outside* the model and that from *inside* the model.

Several fundamental aspects of Skolem’s work are visible in these two results, the Löwenheim–Skolem theorem and the Skolem paradox. Skolem had realized, long before anyone else, that mathematical theories nearly always have several different models. He argued that there are axiom systems, and one can prove theorems in these settings, but what is meant by the objects that obey these rules will generally vary from case to case. From this he drew the radical conclusion that attempts to build mathematics on axiomatic theories were unlikely to succeed (although nowadays, of course, mathematics built on axiomatic foundations has become overwhelmingly successful).

Skolem’s insistence on first-order theories, in which variables may range only over elements, not subsets, was one that his contemporaries took time to accept. But that viewpoint, and the great clarity that comes with it, is today the overwhelmingly dominant one. Skolem insisted that the only possible logic to use in any investigation of the foundations of mathematics was FIRST-ORDER LOGIC [IV.22 §3.2], and that second-order theories were impermissible in the foundations, precisely because second-order theories allowed the axioms to refer to sets, but the nature of sets was, in his view, one of the topics to be elucidated. Skolem also felt that, while one can talk of individual objects, talk of *all* objects of a certain kind can be problematic if it is too informal. Indeed, a generation earlier mathematicians had encountered the paradoxes of naive set theory, where loose talk about all sets of certain kinds causes real difficulties: for example, Russell’s paradox of the set of all sets that are not members of themselves (if it is a member of itself, then it is not, but if it is not, then it is).

Skolem’s work is also characterized by a distrust of the concept of infinity and a preference for finitistic reasoning. He was an early advocate of PRIMITIVE RECURSION [II.4 §3.2.1], which deals with the theory of what are called computable functions, as a way of avoiding paradoxes concerning the infinite.

Further Reading

Fenstadt, J. E., ed. 1970. *Thoralf Skolem: Selected Works in Logic*. Oslo: Universitetsforlaget.

Jeremy Gray

VI.82 Srinivasa Ramanujan

b. Erode, India, 1887; d. Madras (now Chennai), India, 1920
Partitions; modular forms; mock theta functions

Ramanujan, a self-taught Indian genius, made monumental contributions to mathematics that set the stage for many of the breakthroughs in number theory in the twentieth century. He worked on analytic number theory, as well as on ELLIPTIC FUNCTIONS [V.34], hypergeometric series, and the theory of CONTINUED FRACTIONS [III.22]. Much of this work was carried out together with his friend, benefactor, and collaborator G. H. HARDY [VI.73].

Hardy and Ramanujan founded the powerful “circle method” in their remarkable paper that gave an exact formula for $p(n)$, the number of integer partitions of n . Ramanujan independently discovered the two identities that came to be known as the Rogers–Ramanujan identities:

$$1 + \sum_{n=1}^{\infty} \frac{q^{n^2}}{(1-q)(1-q^2) \cdots (1-q^n)} = \prod_{n=0}^{\infty} \frac{1}{(1-q^{5n+1})(1-q^{5n+4})},$$

$$1 + \sum_{n=1}^{\infty} \frac{q^{n^2+n}}{(1-q)(1-q^2) \cdots (1-q^n)} = \prod_{n=0}^{\infty} \frac{1}{(1-q^{5n+2})(1-q^{5n+3})}.$$

These have applications ranging from LIE THEORY [III.50] to statistical physics. The importance of these identities relates to the fact that the GENERATING FUNCTION [IV.18 §§2.4, 3] for $p(n)$ is

$$\prod_{n=0}^{\infty} \frac{1}{1-q^n}.$$

Thus, for example, the second identity asserts that the number of partitions of n into parts all of which are 2 or 3 mod 5 is equal to the number of partitions into distinct parts, all greater than 1, in which no two parts are consecutive integers.

In his work on $p(n)$, Ramanujan discovered and proved many divisibility properties, e.g., that 5 always divides $p(5n+4)$ and that 7 always divides $p(7n+6)$. His conjectures on these divisibility properties inspired the development of extensive methods in MODULAR FORMS [III.61], and his last conjecture was finally settled in 1969 by Oliver Atkin.

All Ramanujan's studies involving $p(n)$ concerned the modular form

$$\eta(w) = q^{1/24} \prod_{n=1}^{\infty} (1 - q^n), \quad \text{where } q = e^{2\pi i w}.$$

The relevance of this is that $q^{1/24}/\eta(w)$ is the generating function for $p(n)$. Of special interest to Ramanujan was the arithmetic function $\tau(n)$, defined by the 24th power of $\eta(w)$: namely,

$$\sum_{n=1}^{\infty} \tau(n) q^n = q \prod_{n=1}^{\infty} (1 - q^n)^{24}.$$

Ramanujan conjectured that $|\tau(p)| < 2p^{11/2}$ for every prime p . The study of this problem led to deep and extensive work on modular forms by H. Petersson, R. Rankin, and others. Eventually, the conjecture was proved by P. Deligne, who received the Fields Medal for his achievement in 1978.

The full story of Ramanujan's life makes his achievements all the more amazing. As a child he was mathematically precocious. In high school he won prizes in mathematics. On the basis of his high school record, he won a scholarship to the Government College in Kumbakonam in 1904. At about this time, Ramanujan came into contact with the book *A Synopsis of Elementary Results in Pure and Applied Mathematics* by G. S. Carr. This rather eccentric book is essentially a huge collection of formulas and theorems compiled for students preparing for the celebrated Mathematical Tripos examination at Cambridge. This book fascinated Ramanujan, who became obsessed with mathematics. In college, he neglected his other subjects and gave his all to mathematics. Consequently, he failed some subjects and lost his scholarship. By 1913, Ramanujan seemed destined for obscurity—he was now a mere clerk in the Madras Port Trust. Friends encouraged him to write to English mathematicians about his mathematical discoveries. Eventually he wrote to G. H. Hardy, who was able to discern that Ramanujan was a truly extraordinary mathematician.

Hardy arranged for Ramanujan to travel to England, and between 1914 and 1918 the two of them produced the groundbreaking work described above.

In 1918, Ramanujan became ill with a sickness diagnosed as tuberculosis. He convalesced in England for a year. His health improved a little in 1919 and he was able to return to India. Unfortunately, his health worsened after his return, and he died in 1920. During this last year in India he penned the pages now known as *Ramanujan's Lost Notebook* and therein laid the foundations of the theory of mock theta functions, a class of

functions similar to but more general than the classical theta functions.

Further Reading

- Berndt, B. 1985–98. *Ramanujan's Notebooks*. New York: Springer.
 Kanigel, R. 1991. *The Man Who Knew Infinity*. New York: Scribners.

George Andrews

VI.83 Richard Courant

b. Lublinitz, Silesia (then part of Germany, now Poland), 1888;
 d. New York, 1972

*Mathematical physics; partial differential equations;
 minimal surfaces; compressible flow; shock waves*

The long and eventful life of Courant was full of high achievements: in mathematical research, the applications of mathematics, as a teacher of many future mathematicians, as a writer of superb books on mathematics, and as an organizer and administrator of large institutions. The fact that Courant—an outsider in his native Germany and a refugee in the United States—could accomplish these things is a testament to his personality as well as to his scientific outlook.

Born in Lublinitz, Courant completed his high school training in Breslau, living on his own and supporting himself by tutoring. His older Breslau friends, Hellinger and Toeplitz, went on to Göttingen, then the mecca of mathematics, and in due course Courant followed them. There he was taken on as an assistant to HILBERT [VI.63], and he began a close friendship with Harald Bohr, which was later extended to Harald's brother Niels.

Under Hilbert's direction, Courant wrote his dissertation on the use of DIRICHLET'S PRINCIPLE [IV.12 §3.5] (on minimizing energy) for constructing conformal maps. Courant also used Dirichlet's principle in several further mathematical studies.

During World War I Courant was drafted into the army as an officer; he fought on the western front and was seriously wounded. After returning to academic life he turned his energies to mathematics and proved some remarkable results: an isoperimetric inequality for the lowest frequency of a vibrating membrane; and the Courant max–min principle for the EIGENVALUES [I.3 §4.3] of a SELF-ADJOINT OPERATOR [III.52 §3.2], so useful in studying the distribution of eigenvalues of the operators of mathematical physics.

In 1920 Courant was named as KLEIN's [VI.57] successor as professor in Göttingen; the appointment was pushed through by Klein and Hilbert, who saw, correctly, that he shared their vision of the relationship between mathematics and science, that he would strike a balance between research and education, and that he had the administrative energy and wisdom to push his mission to fruition.

Courant formed a close friendship with the publisher Ferdinand Springer. One of the fruits of this relationship was the famous "Grundlehren" series of monographs, affectionately known as the "Yellow Peril." The third volume in this series is Courant's exposition of RIEMANN's [VI.49] geometric view of the theory of analytic functions, combined with Hurwitz's lectures on ELLIPTIC FUNCTIONS [V.34]. In 1924 the first volume of Courant-Hilbert on *Mathematical Physics* appeared; it contained, presciently, much of the mathematics needed for Schrödinger's version of quantum mechanics. His influential calculus book appeared in 1927. His research did not languish; in 1928 he published, jointly with his students Friedrichs and Lewy, the basic paper on the difference equations of mathematical physics.

Under Courant's leadership, Göttingen, where the lively international atmosphere had been destroyed by World War I, became once again an important center for mathematics, as well as physics: the list of visitors reads like a Who's Who of mathematics. This was totally shattered when Hitler took over the government: Jewish professors, Courant among the first, were dismissed unceremoniously and had to flee or face annihilation. Courant and his family found refuge in New York, where he was invited to build a Graduate School of Mathematics at New York University (NYU). Without any foundation to build on, Courant succeeded in this task, with the help of his former student Friedrichs and of the American James Stoker, who shared Courant's scientific ideals. Courant found New York a reservoir of talent, and attracted students such as Max Shiffman, and later Harold Grad, Joe Keller, Martin Kruskal, Cathleen Morawetz, Louis Nirenberg, and others, including the writer of this article.

In 1936, in a burst of creativity, Courant obtained basic results about MINIMAL SURFACES [III.96 §3.1] using Dirichlet's principle. In 1937 he finished the second volume of Courant-Hilbert. The immensely successful popular book he wrote jointly with Herb Robbins, *What Is Mathematics?*, appeared in 1940. In 1942 when federal financing for scientific research became

available, Courant's group embarked on an ambitious study of supersonic flow and shock waves.

Federal support did not stop after the war; this enabled Courant to vastly expand the scale of research and graduate instruction at NYU. The research combined, at a high intellectual level, theoretical mathematics with applications such as fluid dynamics, statistical mechanics, the theory of elasticity, meteorology, the numerical solution of partial differential equations, and other topics. Nothing like this had been attempted before at a university in the United States. The institute created by Courant, eventually named after him, is flourishing today and has served as a model for other centers around the world.

Courant hated the Nazis, but did not condemn all Germans; after the war he helped to rebuild mathematics in Germany and was instrumental in inviting talented young German mathematicians and physicists to the United States.

Courant received much help from friends of his youth, many of whom became leaders in their fields, as well as from science administrators in government and industry who admired his vision of mathematics and the gallant spirit that was demonstrated by his willingness to fight against seemingly insuperable odds.

Further Reading

Reid, C. 1976. *Courant in Göttingen and New York: The Story of an Improbable Mathematician*. New York: Springer.

Peter D. Lax

VI.84 Stefan Banach

b. Kraków, Poland, 1892; d. Lwów, Poland, 1945
Functional analysis; real analysis; measure theory; orthogonal series; set theory; topology

Banach was the son of Katarzyna Banach and Stefan Greczek. As his parents were unmarried and his mother was too poor to support her son, he was brought up mainly in Kraków by a foster mother, Franciszka Płowa.

After graduating from high school in 1910, Banach enrolled at the Lwów Polytechnic in the Faculty of Engineering. Two years after his studies were interrupted by the outbreak of World War I, Banach returned to Kraków, where on a summer evening in 1916 he was "discovered" by Hugo Steinhaus, who overheard the words "Lebesgue integral" and brought him to Lwów.

Steinhaus considered this event as his “greatest mathematical discovery.” It was also through Steinhaus that Banach met his future wife, Łucja Braus, whom he married in 1920.

In the same year Professor Antoni Łomnicki engaged Banach as his assistant at the Lwów Polytechnic, even though Banach had not yet finished his studies. This was the beginning of the meteoric rise of Banach’s scientific career.

In June 1920 Banach defended his doctoral dissertation, “On operations on abstract sets and their application to integral equations,” at the Jan Kazimierz University in Lwów. His dissertation was written in Polish and published in 1922 in French. In his thesis Banach introduced the concept of complete normed linear spaces, which are today known as BANACH SPACES [III.64] (the name was proposed by Fréchet in 1928). The theory combined the contributions of RIESZ [VI.74], Volterra, FREDHOLM [VI.66], Lévy, and HILBERT [VI.63] on concrete spaces and on integral equations into a general theory. Banach’s dissertation could be viewed as the birth of *functional analysis*, since Banach spaces are one of its central objects of study.

On April 17, 1922, the Jan Kazimierz University in Lwów awarded Banach his habilitation (a degree allowing him to teach at the university), after which he was appointed Docent in Mathematics. On July 22, 1922, he became a professor of the university (and a full professor from 1927). Banach achieved great research results and became an authority in functional analysis and MEASURE THEORY [III.57]. During the academic year 1924–25 Banach was on sabbatical leave in Paris, where he met LEBESGUE [VI.72], who became a lifelong friend.

In Lwów a group of talented young mathematicians around Banach and Steinhaus soon became the Lwów School of Mathematics and started the journal *Studia Mathematica* in 1929. Among the members of this school were S. Mazur, S. Ulam, W. Orlicz, J. P. Schauder, H. Auerbach, M. Kac, S. Kaczmarz, S. Ruziewicz, and W. Nikliborc. Banach also collaborated with Steinhaus, Saks, and Kuratowski. Many of these mathematicians were later killed by the Germans during the occupation of Poland.

In 1932, Banach’s famous book *Theory of Linear Operations* appeared in French (a Polish version was published the year before) as part of a new series of mathematical monographs, of which he was one of the founders. This was the first monograph on functional analysis as an independent discipline, and it was the

culmination of more than a decade of intense activity by Banach and others.

Banach and the mathematicians around him liked to discuss mathematics in the Café Szkocka (“Scottish café”). This unconventional way of doing mathematics made the atmosphere of Lwów unique—it is one of the rare cases in mathematics of genuine teamwork among a large group. Turowicz and Ulam noted that (see Kaluza 1996, pp. 62, 74):

Banach liked to spend most of his days in a café. He liked the noise and the music. They did not prevent him from concentrating and thinking. It was difficult to outlast or outdrink Banach during these sessions. Problems posed right there were discussed, often with no solution evident even after several hours of thinking. The next day Banach was likely to appear with several small sheets of paper containing outlines of proofs he had completed.

One day in 1935, Banach proposed that the open problems should be collected in a notebook. This notebook later became famous under the name “The Scottish Book.” In the years 1935–41 over 190 problems from various branches of mathematical analysis were proposed in this notebook, and the collection was published in English in 1957 by Ulam. A version with commentaries was published in 1981 by Birkhäuser as *The Scottish Book, Mathematics from the Scottish Café* (edited by R. D. Mauldin).

Banach was also the author of the books *Mechanics* (in two volumes, 1929 and 1930; English translation in 1951), *Differential and Integral Calculus* (in two volumes, 1929 and 1930, with several editions in Polish), *Introduction to the Theory of Real Functions* (in two volumes, written by Banach before the war, although only the first volume remains), and ten textbooks (jointly written with Stożek and SIERPIŃSKI [VI.77]) for primary and secondary schools on arithmetic, geometry, and algebra (published in the years 1930–36 and reprinted in 1944–47).

Banach’s famous discoveries in functional analysis had three important steps. First, he considered abstract linear spaces, where functions are treated like points or vectors, sets of functions as function spaces, and operations on functions as operators. Second, he introduced the *norm* $\| \cdot \|$ of a mathematical object, that is, a quantity that in some (possibly abstract) sense describes the length, size, or extent of the object. The distance between two abstract elements x and y is then given naturally by $d(x, y) = \|x - y\|$. The third important step was to introduce the notion of “complete-

ness” for these spaces. In such general spaces (*Banach spaces*) he was able to prove several fundamental theorems, like the uniform boundedness principle, the open mapping theorem, and the closed graph theorem. What these results say, roughly speaking, is that in a Banach space we cannot have bad (pathological) behavior everywhere—there is always some part of the space where our linear map or other object is well-behaved.

Names like Banach space, Banach algebra, Banach lattice, Banach manifold, Banach measure, Hahn–Banach theorem, Banach fixed-point theorem, Banach–Mazur game, Banach–Mazur distance between isomorphic spaces, Banach limits, Banach–Saks property, Banach–Alaoglu theorem, and the Banach–Tarski paradox on a decomposition of sets into congruent parts show how wide his influence has been. Banach also introduced the notions of DUAL SPACE [III.19], dual operator, the general concepts of weak and weak-star convergence, and he used all of these notions in linear operator equations.

In 1936, Banach delivered a one-hour plenary address at the International Congress of Mathematicians in Oslo, where he described the work of the whole Lwów school. In 1937 Norbert Wiener tried to lure him to the United States. In 1939 he was elected president of the Polish Mathematical Society and was awarded a Grand Prize of the Polish Academy of Knowledge. Banach spent the war years in Lwów. During the years 1940–41 and 1944–45 he was the Dean of the Faculty of Science at the renamed Iwan Franko State University. In the period 1941–44 Lwów was occupied by the German army. During this period Banach was saved from almost certain death by the action of Rudolf Weigel, a “Schindleresque” factory owner and inventor of the typhus vaccine, who gave him employment at his Bacteriological Institute as a louse feeder. After the war, he accepted a chair at the Jagiellonian University. He died on August 31, 1945, in Lwów of lung cancer at the age of fifty-three.

The complete list of Banach’s publications comprises fifty-eight items, and they were reprinted in Banach’s *Collected Works* (published in two volumes in 1967 and 1979). Banach said, “Mathematics is the most beautiful and most powerful creation of the human spirit. Mathematics is as old as Man.” Banach is considered a national hero in Poland, as a great scientist and a major figure in the great flowering of Polish scientific life in the independent Poland of the interwar years.

Further Reading

Banach, S. 1967, 1996. *Oeuvres*, two volumes. Warsaw: PWN.
Kaluza, R. 1996. *The Life of Stefan Banach*. Basel: Birkhäuser.

Lech Maligranda

VI.85 Norbert Wiener

b. Columbia, Missouri, 1894; d. Stockholm, 1964
Stochastic processes; applications to electrical engineering and physiology; harmonic analysis; cybernetics

Wiener was just eighteen years old when, in 1913, he was awarded a Ph.D. in logic while studying under Josiah Royce at Harvard University. Afterward, he studied with, among others, RUSSELL [VI.71] and HARDY [VI.73] in Cambridge and HILBERT [VI.63] in Göttingen. After doing work on ballistics for the military during World War II, he was appointed instructor of mathematics at the fledgling Massachusetts Institute of Technology in Cambridge, MA, where he remained for the rest of his career.

Wiener was in many respects a nonconformist, certainly scientifically and mathematically, but also socially, culturally, politically, and philosophically. He was a precocious child and his home education by his father (a noted linguist and Harvard professor), along with his Jewish background in a society still stricken by anti-Semitism, made his nonconformism almost inevitable. Garrett Birkhoff, the son of GEORGE BIRKHOFF [VI.78], said the following in 1977:

Wiener was notable as one of the few Americans of his time who was outstanding in both pure mathematics and its applications. How much of this can be attributed to his varied and cosmopolitan early background, and how much to his continuing contacts with non-mathematicians ... it is hard to say.

During a period in which American mathematics was largely self-sufficient and was still in a phase in which interdisciplinary approaches were generally ignored, Wiener was reaching out to European mathematics and collaborating with engineers such as Vannevar Bush.

This attitude also affected his choices of research topics, even within pure mathematics: he worked on whatever took his fancy. In a talk in 1938, George Birkhoff described Wiener’s work on Tauberian theorems as an example of “exercising talent for free invention,” contrasting this with the typically American approach: “mathematics as serious business.”

Wiener's way of connecting pure and applied mathematics did not follow the usual path of taking old problems of applied mathematics (such as in classical mechanics and electrical engineering) and tackling them with new, and rigorously sharpened, mathematical tools. Rather the opposite: Wiener used some of the newest, and much debated, results of pure mathematics—such as the LEBESGUE INTEGRAL [III.57], Fourier transformations in the complex domain, and STOCHASTIC PROCESSES [IV.24]—and connected them to several of the newest physical, technological, and biological problems. The types of problem he attacked included those of BROWNIAN MOTION [IV.24], quantum mechanics, radio astronomy, anti-aircraft fire control, noise filtration in radar, the nervous system, and the theory of automata.

Of Wiener's many analytical results that make connections between very different domains we give only one as an example. Around 1931 Wiener discussed the following (Lebesgue) integral equation with the German mathematical astrophysicist Eberhard Hopf:

$$f(t) = \int_0^\infty W(t - \tau)f(\tau) d\tau.$$

The solution for unknown $f(t)$, which was found with the help of a new and very important factorization technique, and which was dependent on the analytical behavior of the FOURIER TRANSFORMS [III.27] of the functions involved, could be connected to radiative equilibrium in stars. When t is interpreted as time, equations of this kind can be seen to describe causality: the transition from the influencing "past" to the indeterminate "future." A decade later the Wiener-Hopf equations would be connected to Wiener's theories of prediction and filtering.

Wiener's discussion of such disparate fields of application could not fail to invoke philosophically relevant notions such as causality, information (Wiener is considered together with Claude Shannon as the founder of the modern concept of information), control, feedback, and finally the wide-ranging theory of "cybernetics." Cybernetics (literally, the art of steering) can be retrospectively connected to earlier discussions in Greek antiquity (Plato), to James Watt's centrifugal governor, and to Ampère's philosophical writings. Wiener's broad outlook resulted from his collaboration with colleagues from very different domains: mathematical (R.E.A.C. Paley), physical (Hopf), technical (Julian Bigelow, Bush), and physiological (Arturo Rosenblueth). However, this outlook left him vulnerable to criticism and to philosophical and political misinterpretation.

The prominent mathematician Hans Freudenthal was a sharp-tongued critic of Wiener's epoch-making book of 1948, *Cybernetics or the Control and Communication in the Animal and the Machine*, claiming that it "shows there is not much to be reported" and that it "has contributed to spreading mistaken ideas of what mathematics really means," although even he had to admit that the book "earned Wiener the greater part of his public renown" and that its "mathematical readers were more fascinated by the richness of its ideas than by its shortcomings."

During the period of the Nazi threat Wiener helped refugees from Europe to settle in the United States, while after World War II he cautioned against the repetition of mistakes such as the boycott of German science in the aftermath of World War I. Wiener warned against the arms race and the misuse of technological developments in the postwar world. Having resigned from the National Academy of Sciences in 1941 because of its alleged bureaucracy and complacency, Wiener nevertheless accepted, shortly before his death in 1964 while traveling, the National Medal of Science from President Johnson.

Further Reading

Masani, P. R. 1990. *Norbert Wiener 1894-1964*. Basel: Birkhäuser.

Reinhard Siegmund-Schultze

VI.86 Emil Artin

b. Vienna, 1898; d. Hamburg, Germany, 1962
Number theory; algebra; theory of braids

Born in fin de siècle Vienna to an art dealer father and opera singer mother, Artin was influenced throughout his life by the rich cultural atmosphere of the late Hapsburg Empire. He was, as the algebraist Richard Brauer described him, as much artist as mathematician. After his first semester at the University of Vienna in 1916, Artin was drafted into the Austrian Army, in which he served until the end of World War I. In 1919 he enrolled at the University of Leipzig, and completed his doctorate under the direction of Gustav Herglotz in only two years.

Artin spent the academic year 1921-22 at the mathematically vibrant University of Göttingen, and then moved to the recently opened University of Hamburg. He achieved the rank of full professor in 1926. While

at Hamburg, Artin oversaw the work of eleven doctoral students, including Max Zorn and Hans Zassenhaus. Artin's years at Hamburg were among the most productive of his life.

Artin's work in CLASS FIELD THEORY [V.31], the subject closest to his heart, led him to a solution of Hilbert's ninth problem: a proof of the most general law of reciprocity. The aim was to generalize Gauss's law of quadratic reciprocity and the higher reciprocity laws. Teiji Takagi's fundamental results on class field theory had appeared when Artin was a student. Using Takagi's theory, N. G. Chebotaryov's 1922 proof of the density theorem (conjectured by FROBENIUS [VI.58] in 1880), and his own theory of L -FUNCTIONS [III.49], Artin established his general law of reciprocity in 1927. Artin's theorem not only provided the final form of the classical question on reciprocity but it also formed the central result of class field theory. Both Artin's result and his tools, particularly his L -functions, proved important. Artin posed a conjecture about his L -functions that remains unanswered today. Questions in non-Abelian class field theory also remain open.

In 1926–27, Artin and Otto Schreier developed the theory of formally real closed fields: fields with the property that -1 cannot be expressed as the sum of two squares (an example being the real numbers). This work formed the basis of Artin's solution of Hilbert's seventeenth problem concerning rational functions.

Artin extended Wedderburn's theory of algebras ("hypercomplex numbers") to noncommutative rings with chain conditions in 1928. Indeed, the class of such rings called "Artinian rings" is named in his honor.

In 1929 Artin married one of his students, Natalie Jasny. Natalie's Jewish background and Artin's personal sense of justice prompted them to leave Germany in 1937. They emigrated to America, where Artin spent a year at Notre Dame University before moving to a permanent position at Indiana University. Artin's lectures at Notre Dame led to his influential text *Galois Theory* (1942), which reflected his quest for simplification and his desire to unite different research trends.

At Indiana, Artin began a collaboration with George Whaples of the University of Pennsylvania and introduced the concept of a valuation vector, a notion closely related to the concept of an idèle introduced by Claude Chevalley. This work seemed to revitalize Artin's mathematical research, and, after something of a hiatus in his written work, he began to publish regularly again.

In 1946, Artin moved to Princeton University. While there, Artin oversaw eighteen of his thirty-one Ph.D.

students, including John Tate and Serge Lang. He also returned to his work in the theory of BRAIDS [III.4], a topic that relates questions in topology and group theory. His introduction to the theory of braids that appeared in *American Scientist* in 1950 reveals Artin's prowess as a master expositor.

Further Reading

Brauer, R. 1967. Emil Artin. *Bulletin of the American Mathematical Society* 73:27–43.

Della Fenster

VI.87 Alfred Tarski

b. Warsaw, 1901; d. Berkeley, California, 1983

Symbolic logic; metamathematics; set theory; semantics; model theory; algebras of logic; universal algebra; axiomatic geometry

Tarski matured during Poland's renaissance in mathematics and philosophy in the remarkable interwar period of Polish independence. His teachers at the University of Warsaw included Stanisław Leśniewski and Jan Łukasiewicz in logic, SIERPIŃSKI [VI.77] in set theory, and Stefan Mazurkiewicz and Kazimierz Kuratowski in topology. In his thesis Tarski solved a core problem in Leśniewski's idiosyncratic system for the foundation of mathematics, but afterward he focused on set theory and more mainstream mathematical logic. Almost immediately he obtained the spectacular BANACH-TARSKI PARADOX [V.3] (that it is possible to dissect a solid sphere into a finite number of pieces that may then be reassembled to form *two* spheres of the same radius as the original one) in collaboration with BANACH [VI.84].

Encouraged by his professors, he changed his original surname, Teitelbaum, to Tarski just before receiving his Ph.D. in 1924, because a Jewish name was a professional handicap. This accorded with Tarski's strong identification with Polish nationalism and his belief that assimilation was a rational solution to the Jewish question.

By 1930, Tarski had established one of his most important results: the completeness and decidability of formal systems of the algebra of real numbers and of Euclidean geometry axiomatized within first-order logic (see LOGIC AND MODEL THEORY [IV.23 §4]). In the following years Tarski concentrated on fundamental conceptual developments in metamathematics and the semantics of formalized languages. In contrast with

HILBERT [VI.63], who called for the execution of his metamathematical consistency program by the most restricted means possible, Tarski was open to the use of any mathematical methods, including all those of set theory. His main conceptual contribution was to provide a theory of truth for formalized languages, in which he laid down a novel criterion—called the T-scheme—for an adequate definition of truth for such a language, and showed how it can be met by a set-theoretical definition within a metalanguage, while it cannot be defined within the language itself.

Though Tarski's preeminence in Polish logic was widely acknowledged, he never succeeded in obtaining a chair in his country of birth, partly because of the paucity of positions, and partly as a result of anti-Semitism, notwithstanding his change of name. Made a Docent at the University of Warsaw as soon as he had finished his Ph.D., his position was later raised to that of Adjunct Professor. Neither post paid a living wage, and so, in order to make ends meet, Tarski also taught in a Gymnasium (high school) throughout the 1930s. Because he did not hold a chair, he could not be designated the official director of the dissertation of his first student, Andrzej Mostowski; instead Kuratowski assumed that role.

An invitation to attend a Unity of Science meeting (an offshoot of the Vienna Circle) at Harvard brought Tarski to the United States two weeks before the Nazi invasion of Poland on September 1, 1939. Given his Jewish origins, this probably saved his life, but the war separated him from his family. (His wife and immediate family survived the war, but most of the rest of his family perished in the Holocaust.) In the United States he was granted a permanent nonquota visa within months, but only temporary positions were available to him during the period 1939–42. Finally, he succeeded in obtaining a position as Lecturer in the Department of Mathematics of the University of California at Berkeley. There Tarski's manifest excellence was soon recognized and he rose rapidly to the position of Full Professor by 1946. In the following decade, through his charismatic teaching and zealous campaigning for additional appointments in the field, he built a program in logic and the foundations of mathematics that made Berkeley a mecca for logicians from all over the world for years to come.

It was not until 1939 that Tarski wrote up his decision procedure for algebra and geometry for publication; it was slated to appear as a monograph for a Parisian publisher, but that was aborted following

the invasion of France by Germany in 1940. A revised version with full details was finally prepared with the assistance of J.C.C. McKinsey as a RAND Corporation report in 1948; it only became publicly available a few years later through the University of California Press. This work then became paradigmatic for the applications of model theory to algebra in which the Tarski school led the way; the subject has continued to be one of the most important parts of mathematical logic to this day. At Berkeley during the postwar period Tarski also promoted substantial developments along several different lines: algebraic logic, the axiomatization of set theory and the significance of LARGE CARDINAL [IV.22 §6] assumptions for mathematical problems, and the axiomatization of geometry. Above all, the importance of Tarski's work lay in opening the field of logic to the unrestricted use of set-theoretical methods, combined with a constant attention to rigorous and proper conceptual development.

Further Reading

- Feferman, A. B., and S. Feferman. 2004. *Alfred Tarski. Life and Logic*. New York: Cambridge University Press.
 Givant, S. 1999. Unifying threads in Alfred Tarski's work. *Mathematics Intelligencer* 13(3):16–32.
 Tarski, A. 1986. *Collected Papers*, four volumes. Basel: Birkhäuser.

*Anita Burdman Feferman and
Solomon Feferman*

VI.88 Andrei Nikolaevich Kolmogorov

b. Tambov, Russia, 1903; d. Moscow, 1987
Analysis; probability; statistics; algorithms; turbulence

Kolmogorov was one of the greatest mathematicians of the twentieth century. His work was distinguished both by its great depth and power, and by its breadth: he made important contributions to several different areas. He is most famous for his work on probability theory, and is widely regarded as having been the greatest probabilist ever.

Kolmogorov's mother, Mariya Yakovlena Kolmogorova, died in childbirth; his father, Nikolai Matveevich Kataev, an agronomist, worked for the Ministry of Agriculture after the Revolution and died in the Denikin offensive in the Civil War in 1919. Kolmogorov was brought up by his mother's sister Vera, whom he regarded as his mother and who lived to see her adopted son's success.

After his childhood in Tunoshna, near Yaroslavl on the Volga, Kolmogorov became a student of mathematics at Moscow University in 1920. His teachers included Aleksandrov, Lusin, Urysohn, and Stepanov. Kolmogorov's first work, published in 1923 when he was still nineteen, gave an example of a (Lebesgue integrable) function whose FOURIER SERIES [III.27] diverges almost everywhere. (This is in contrast to the classical theorems giving regularity conditions on a function that are sufficient for its Fourier series to converge to it.) This famous and unexpected result made him a celebrity, all the more so when in 1925 he sharpened "almost everywhere" to "everywhere."

Kolmogorov became a postgraduate student in 1925, studying under Lusin. Also in 1925, he published his first work on probability theory, in collaboration with Alexander Yakovlevich Khinchin (Khintchine, Hincin), on the "three series theorem." This classical result gives a necessary and sufficient condition for the convergence of a random series with independent terms, namely the convergence of three nonrandom series. The paper also contains the Kolmogorov inequality on maxima of independent sums. By the time of his doctorate in 1929, Kolmogorov had written eighteen mathematical papers: on analysis, on probability, and on intuitionist logic, an indication of his lifelong interest in the foundations of mathematics. He became a professor at Moscow University in 1931.

Also in 1931, Kolmogorov published his famous paper on analytic methods in probability theory. This deals with Markov processes in continuous time, with the state space continuous or discrete (in which case one speaks of a Markov chain). The Chapman-Kolmogorov equations, and the Kolmogorov forward and backward differential equations, date from this paper. Diffusions are also treated, developing earlier work by Bachelier.

The whole subject of modern probability theory was given a firm foundation by Kolmogorov's epoch-making monograph *Grundbegriffe der Wahrscheinlichkeitsrechnung* of 1933 (later translated as *Foundations of Probability Theory*). Before this time, probability had lacked a rigorous mathematical foundation, and indeed some authors had believed that it was impossible to provide one. However, the relevant mathematical theory, MEASURE THEORY [III.57], had been introduced by LEBESGUE [VI.72] in 1902, in connection with his theory of the integral. Measure theory also provides a firm foundation for the mathematics of length, area, and volume.

By the 1930s, the subject had been freed from its origins in Euclidean space. Kolmogorov treated probability simply as a measure of total mass 1, events as measurable sets, RANDOM VARIABLES [III.73 §4] as measurable functions, etc. The decisive technical innovation was his treatment of conditioning, which used the then-recent Radon-Nikodým theorem (whereby conditional expectations became Radon-Nikodým derivatives). The *Grundbegriffe* also contains two further key results. The first is the Daniell-Kolmogorov theorem, basic to the definition of a STOCHASTIC PROCESS [IV.24]. The second is Kolmogorov's STRONG LAW OF LARGE NUMBERS [III.73 §4]. When we repeatedly toss a fair coin, we expect the observed frequency of heads to tend to the expected frequency, a half. Some restriction is needed to make precise mathematical sense out of this intuition. It was known before Kolmogorov that the qualification needed here is that convergence takes place with probability 1 ("almost surely," or "a.s."). Kolmogorov generalized this result from coin tossing to repeated replication of any random experiment. One needs the expected value (often called the mean) to exist, in the technical sense of measure theory. Then the average value in a sample, the *sample mean*, converges to the expectation, the *population mean*, with probability 1.

Further work by Kolmogorov on probability theory followed in the 1930s and 1940s. He worked on limit theorems, on infinite divisibility, on the Kolmogorov-Petrovskii-Piscunov equation governing the wave of advance of an advantageous gene, and on linear prediction of stationary stochastic processes. This application, which led to the "Kolmogorov-Wiener filter," was motivated by wartime applications to fire control problems.

This last work led Kolmogorov naturally to path-breaking work on turbulence in 1941, including the Kolmogorov "two-thirds power" law. This work has been profoundly important subsequently, as the problem of understanding turbulence is a central one in fluid dynamics.

Motivated by questions of the stability of the solar system, and related DYNAMICAL SYSTEMS [IV.14], Kolmogorov published in 1954 his work on mechanics and invariant tori, work that developed into the subject of "KAM theory" (for Kolmogorov, Arnold, and Moser).

Kolmogorov's axiomatization of probability theory can be regarded as a solution of (part of) Hilbert's sixth problem, to put probability and mechanics onto a rigorous footing. In 1956 and 1957, Kolmogorov solved

another of Hilbert's problems, the thirteenth. His solution gave a surprising structure theorem, by which a function of many variables can be built up from functions of few variables by means of basic operations. He showed that a continuous function of any number of real variables may be built up by combining (using the operations of addition and of taking a function of a function) a finite number of functions of *only three* real variables. He regarded this work as his most technically difficult accomplishment.

In the 1960s, Kolmogorov's interests shifted to foundational questions: in mathematics, in probability theory, and in INFORMATION THEORY [VII.6] and the theory of algorithms. He introduced the concept now called "Kolmogorov complexity." He gave a new approach to randomness, quite different from that in his earlier work on probability theory. Here, random sequences are identified as *sequences of maximal complexity*. His later work was dominated by his lifelong interest in teaching, and in particular his involvement in special schools for particularly gifted pupils.

Kolmogorov's *Selected Works* comprise three volumes: *Mathematics and Mechanics*, *Probability and Statistics*, and *Information Theory and Algorithms*.

He was widely honored, both within the Soviet Union and outside. He was married, with no children.

Further Reading

- Kendall, D. G. 1990. Obituary, Andrei Nikolaevich Kolmogorov (1903–1987). *Bulletin of the London Mathematical Society* 22(1):31–100.
- Shiryayev, A. N., ed. 2006. *Selected Works of A. N. Kolmogorov*. New York: Springer.
- Shiryayev, A. N., and others. 2000. *Kolmogorov in Perspective*. History of Mathematics, volume 20. London: London Mathematical Society.

Nicholas Bingham

VI.89 Alonzo Church

b. Washington, District of Columbia, 1903; d. Hudson, Ohio, 1995
Logic

Church's career was spent almost entirely at Princeton. Having studied there, he returned, after spells in Harvard, Göttingen, and Amsterdam, to take up a position as an assistant professor in 1929, and rose to become Professor of Philosophy and Mathematics in 1961, a position he held until his retirement in 1967. He then moved to the University of California at Los

Angeles where he was Kent Professor of Philosophy and Professor of Mathematics until he retired (again) in 1990.

Princeton became an important center for logic during the 1930s: VON NEUMANN [VI.91] arrived at the beginning of the decade, GÖDEL [VI.92] visited in 1933 and 1935 before moving there permanently in 1940, and from September 1936 TURING [VI.94] spent two years there as a graduate student, completing his Ph.D. with Church.

In 1936 Church made two profound contributions to the theory of logic. The first, which appeared in a paper entitled "An unsolvable problem in elementary number theory," is what is now known as *Church's thesis*: the proposal to identify the vague intuitive notion of effective calculability with the precise notion of a RECURSIVE FUNCTION [II.4 §3.2.1]. It quickly transpired that Church's definition of a recursive function was equivalent to Turing's definition of computable functions. At the end of 1936, Turing, who had been working with analogous ideas in an entirely different way, published his famous paper "On computable numbers," which contained the result that every function that is naturally regarded as computable is computable by a TURING MACHINES [IV.20 §1.1]. Church's thesis is therefore often known as the Church–Turing thesis.

The second of Church's contributions is what is now known as *Church's theorem*. In a short paper published in the first issue of *The Journal of Symbolic Logic*, Church showed that it is impossible to decide algorithmically whether statements in arithmetic are true or false. It follows that a general solution to the *Entscheidungsproblem* (decision problem) does not exist; equivalently, first-order logic is undecidable. This result is also known as the Church–Turing theorem since Turing independently (and in the paper referred to above) proved the same result (see THE INSOLUBILITY OF THE HALTING PROBLEM [V.23]). In achieving this result, both Church and Turing were strongly influenced by GÖDEL'S INCOMPLETENESS THEOREM [V.18].

VI.90 William Vallance Douglas Hodge

b. Edinburgh, Scotland, 1903; d. Cambridge, England, 1975
Algebraic geometry; differential geometry; topology

Hodge is famous for his theory of harmonic integrals (or forms), which was described by WEYL [VI.80] as "one of the landmarks of twentieth century mathematics." He was a Scot who spent his early life in Edinburgh but lived for most of his life in Cambridge, where he was

Lowndean Professor of Astronomy and Geometry (an archaic title) from 1936 until 1970.

Hodge's work straddles the area between algebraic geometry, differential geometry, and complex analysis. It can be seen as a natural outgrowth of the theory of RIEMANN SURFACES [III.81] (or algebraic curves) and the work of Lefschetz on the topology of algebraic VARIETIES [IV.4 §7] (of higher dimension). It put algebraic geometry on a modern analytic footing and prepared the ground for the spectacular breakthroughs of the postwar period in the 1950s and 1960s. It also harmonized well with the later interaction with theoretical physics, harking back to the influence of James Clerk Maxwell.

In Riemann surface theory (with one complex dimension), complex structures and real metrics are very closely related and the roots of their relationship can be traced back to the link between the CAUCHY-RIEMANN EQUATIONS [I.3 §5.6] and the LAPLACE OPERATOR [I.3 §5.4]. In higher dimensions this close link disappears and a RIEMANNIAN METRIC [I.3 §6.10] seems alien to complex analysis, but it was Hodge's great insight to see that real analysis could still play a fruitful role.

Following the formalism of electromagnetic theory as developed by Maxwell, he introduced a generalization of the Laplace operator to exterior DIFFERENTIAL FORMS [III.16] (on any Riemannian manifold) and proved the key theorem that the null space of this operator on r -forms ("harmonic" forms) is naturally isomorphic to the r -dimensional COHOMOLOGY [IV.6 §4] H^r . In other words, a harmonic form is uniquely specified by its periods, and all sets of periods occur.

For complex manifolds, provided the metric is suitably compatible with the complex structure (the KÄHLER CONDITION [III.90 §3], always satisfied by algebraic varieties in projective space), this result can be refined. We get a decomposition of H^r into subspaces $H^{p,q}$ with $p + q = r$, with the extreme cases $p = r$, $q = r$ corresponding to holomorphic or anti-holomorphic forms.

This Hodge decomposition has a rich structure and a wealth of applications. One of the most remarkable is the Hodge signature theorem, which (for an even-dimensional algebraic variety) expresses the signature of the intersection matrix of middle-dimensional cycles in terms of the dimensions of the $H^{p,q}$.

Another success of the theory was the characterization of those homology classes of dimension $2n - 2$ (on a complex n -manifold) that arise from algebraic

subvarieties. He conjectured that a similar characterization would work for all dimensions and proved the easy part. The hard part has resisted all subsequent attempts, and is now one of the million-dollar Millennium Problems of the Clay Institute.

The influence of Hodge's theory was enormous. First, in algebraic geometry it integrated many classical results into a modern framework and it acted as a launch pad for the subsequent development of modern sheaf theory by Henri Cartan, Serre, and others. Second, it was the first deep result in global differential geometry and paved the way for what became known as "global analysis." Third, it provided the basis for later developments arising from, or linked to, theoretical physics. These included THE ATIYAH-SINGER INDEX THEOREM [V.2] for elliptic operators, and nonlinear analogues of Hodge theory (the Yang-Mills and the Seiberg-Witten equations), which have played such a key role in the Donaldson theory of four-dimensional manifolds (see DIFFERENTIAL TOPOLOGY [IV.7 §2.5]). More recently, Witten and others have shown how suitable infinite-dimensional versions of Hodge's theory turn up naturally in QUANTUM FIELD THEORY [IV.17 §2.1.4].

Further Reading

Griffiths, P., and J. Harris. 1978. *Principles of Algebraic Geometry*. New York: Wiley.

Sir Michael Atiyah

VI.91 John von Neumann

b. Budapest, 1903; d. Washington, District of Columbia, 1957
Axiomatic set theory; quantum physics; measure theory; ergodic theory; operator theory; algebraic geometry; theory of games; computer engineering; computer science

PUP: 'District of Columbia' spelled out here. Appropriate in the context of spelling out state names?

Raised as a Hungarian Jew in the Austrian Empire, Neumann János Lajos's political outlook was strongly affected by the five-month reign of the communist Béla Kun's regime after World War I. It formed his liberal and democratic political credo (although he did insist on retaining the title of nobility "margittai," acquired by his father in 1913, which he later translated to the German "von"). He was a child prodigy, learning several languages and demonstrating an early enthusiasm for mathematics.

During the early 1920s von Neumann studied mathematics, physics, and chemistry in Berlin and Zürich, and was also enrolled to study mathematics in Budapest

although he never attended any lectures there. He received a diploma in chemical engineering at the ETH Zürich and shortly afterward (in 1926) a doctorate in mathematics at the University of Budapest (his thesis was entitled “The axiomatic deduction of general set theory”). While engineering was considered a respectable profession for a brilliant young man with such wide-ranging interests, the theoretical challenges of mathematics and formal logic drove von Neumann to the more academic environment in Germany, where he immediately received attention from HILBERT [VI.63]. Although the sensible choice, academically speaking, would have been to stay with Hilbert at Göttingen—and he did spend six months there during 1926–27 on a Rockefeller Fellowship—he preferred the pulsating atmosphere of Berlin.

During the following years he published on the axiomatic foundations of set theory, on MEASURE THEORY [III.57], and on the mathematical foundations of quantum mechanics. He also wrote his first paper on game theory (“Zur Theorie der Gesellschaftsspiele,” published in *Mathematische Annalen* in 1928), proving the *minimax theorem* (the theorem that states that every two-person finite zero-sum game has optimal mixed strategies).

In 1927 von Neumann received his habilitation in mathematics from the Philosophical Faculty of Berlin University with a written thesis and a lecture on the foundations of set theory and mathematics, becoming one of the youngest Privatdozenten in the history of the university. At this point he changed his name to the German Johann von Neumann. He gave lecture courses in Hamburg (1929–30) as well as in Berlin, but in 1933, with the Nazi seizure of power, he resigned from his appointment at Berlin. By that time he was already in Princeton, where his visiting status at the university, originally conferred in 1930, was transformed into a tenured position at the newly founded Institute for Advanced Study. He modified his name once again, this time to John von Neumann, receiving U.S. citizenship in 1937.

At Princeton he found a peaceful ivory tower. Much of his important mathematical work stems from that period in the mid 1930s: he published around six journal articles per year (a rate he maintained until his death), as well as several books. The Institute’s environment allowed him to expand his research scope, taking in, among other things, ERGODIC THEORY [V.11], Haar measure, certain spaces of operators on a HILBERT SPACE [III.37] (these spaces are now known as

VON NEUMANN ALGEBRAS [IV.15 §2]), and “continuous geometry.”

Von Neumann was much too politically sensitive to ignore the European crisis that led to World War II. Having begun to investigate turbulent flow beyond the speed of sound in the mid 1930s, he was invited to the Ballistic Research Laboratory in 1937 as an expert on shock waves. Later he acted as a consultant to the Navy and the Air Force. Although he was not in the initial group of Los Alamos scientists, in 1943 he became an advisor to the Manhattan Project, where his mathematical treatment of shock waves became essential, leading to the “implosion lens,” an arrangement of explosives that started the uranium chain reaction.

In parallel with his war-related work, von Neumann pursued his interest in economics, which resulted in a collaboration with Oskar Morgenstern: their groundbreaking book *The Theory of Games and Economic Behavior*, partly based on his 1928 *Mathematische Annalen* paper, appeared in 1944.

In the 1940s von Neumann began to focus on computing as a result of two very different branches of his thinking: namely, the numerical approximation of solutions to otherwise unsolvable problems, and his proficiency in the foundations of mathematics. He had tried to enlist TURING [VI.94] as an assistant at Princeton and he was certainly aware of the importance of Turing’s seminal paper on computable numbers (1936). While Turing discussed an abstract machine in the form of a thought experiment, von Neumann also considered the problems arising from the actual construction of computers, such as those connected with the use of electronic hardware. His training as a mathematician allowed him to focus on the very essentials of computing machinery and avoid baroque designs like the Moore School’s ENIAC (Electronic Numerical Integrator And Computer). In 1945 he defined the essential components for the “Electronic Discrete Variable Computer.” His “First draft of a report on the EDVAC,” which summarized and focused ideas gathered from work on early electronic computers, provided a logical framework for the modern electronic computer, becoming a road map for computer architecture for the ensuing decades. While von Neumann probably did not consider this paper to have the same importance as his mathematical results, today it is considered the birth certificate of modern computers.

Von Neumann quickly recognized that programming computers (or “coding,” as he called it) was likely to be more demanding than building basic hardware. In

essence he considered programming as a new branch of formal logic. In 1947 he coauthored (with Herman Goldstine) a three-part report, "Planning and coding of problems for an electronic computing instrument," in which many insights on the novel and demanding art of software construction were collected together.

Von Neumann's thinking went beyond the restrictions of calculating machines, and allowed him to venture into philosophical questions on the structure of the human brain and cellular automata and the idea of self-reproducing systems—questions that were forerunners to the disciplines now called "artificial intelligence" and "artificial life." Consideration of these questions resulted in a series of lectures published as *The Computer and the Brain* (1958) and a book, *Theory of Self-Reproducing Automata* (1966), both of which appeared posthumously.

In 1954 von Neumann was appointed to the five-member U.S. Atomic Energy Commission and in 1956 he was awarded the Presidential Medal of Freedom by President Eisenhower.

Further Reading

Aspray, W. 1990. *John von Neumann and the Origins of Modern Computing*. Cambridge, MA: MIT Press.

Wolfgang Coy

VI.92 Kurt Gödel

b. Brno, Moravia (now Czech Republic), 1906;
d. Princeton, New Jersey, 1978
Logic; relativity theory

Born in Brno, Moravia, Gödel did his most important work at the University of Vienna. In 1940 he emigrated to the United States, where he accepted an appointment at the Institute for Advanced Study in Princeton.

Considered the greatest mathematical logician of the twentieth century, Gödel is renowned for his proofs of three fundamental results: the semantic COMPLETENESS OF FIRST-ORDER LOGIC [IV.23 §2]; the syntactic INCOMPLETENESS OF FORMAL NUMBER THEORY [V.18]; and the consistency, relative to the AXIOMS OF ZERMELO-FRAENKEL [IV.22 §3.1] set theory, of the AXIOM OF CHOICE [III.1] and the generalized CONTINUUM HYPOTHESIS [IV.22 §5].

Gödel's completeness theorem (1930) is concerned with the following kind of question: how do we know that a statement in group theory, for example, that is

true in every group is actually provable from the axioms of group theory? Gödel showed that in any first-order theory (one in which quantifiers are allowed over elements but not over subsets), any statement true in all models is indeed provable. In an equivalent form, this completeness theorem states that any set of statements that is consistent (that is, from which no contradiction may be derived) has a model—a structure in which all those statements hold.

Gödel's incompleteness theorem (1931) sent shock waves through logic and the philosophy of mathematics. HILBERT [VI.63] had set out a program in which all statements (in number theory, for example) should be derivable from a fixed set of axioms. It was generally believed that such a program was in principle possible, until the incompleteness theorem destroyed that hope.

Gödel's idea was to construct a statement S that, in effect, asserts " S is not provable." A moment's thought shows that such a statement must be both true and unprovable. Gödel's remarkable achievement was to manage to encode such a statement in the language of number theory. His proof applies to such axioms as THE PEANO AXIOMS [III.69] for number theory, and more generally, to any reasonable extension of them (such as the Zermelo-Fraenkel axioms for set theory).

Gödel's second incompleteness theorem represented another blow to the Hilbert program. Suppose that we have a set of axioms T (for example, the Peano axioms) that is consistent. Can we prove that it is consistent? Gödel showed that, if T is consistent, then the statement " T is consistent" (when encoded as a statement of number theory) cannot be proved from T . So " T is consistent" is an explicit example of a true but unprovable statement. Again, this applies when T is the set of Peano axioms, or any reasonable extension thereof (roughly, any extension that allows one to encode into arithmetic statements about provability and the like). As a slogan: "a theory cannot prove its own consistency."

The axiom of choice became highly controversial when Ernst Zermelo used it to prove that every set can be well-ordered, a task that, together with the proof of the continuum hypothesis, Hilbert had listed first among the problems he posed in 1900 to the International Congress of Mathematicians. In 1938 Gödel showed that both the axiom of choice and the generalized continuum hypothesis are consequences of another principle (the axiom of constructibility) that holds in a submodel of any model of Zermelo-Fraenkel set theory. Both are consequently consistent with (not

disprovable from) the Zermelo–Fraenkel axioms. Much later (1963) Paul Cohen showed that both statements are also independent of (not provable from) those axioms.

Apart from logic, Gödel also worked in relativity theory, where he established the existence of models of EINSTEIN’S FIELD EQUATIONS [IV.13] that permit time travel into the past.

Further Reading

Dawson Jr., J. W. 1997. *Logical Dilemmas: The Life and Work of Kurt Gödel*. Natick, MA: A. K. Peters.

John W. Dawson Jr.

VI.93 André Weil

b. Paris, 1906; d. Princeton, New Jersey, 1998
Algebraic geometry; number theory

André Weil was one of the most influential mathematicians of the twentieth century. His influence is due both to his original contributions to a remarkably broad spectrum of mathematical theories, and to the mark he left on mathematical practice and style, through some of his own works as well as through the BOURBAKI [VI.96] group, of which he was one of the principal founders.

Weil, as well as his sister, the philosopher, political activist, and religious thinker Simone Weil, received an excellent education. Both were brilliant students, very widely read, with a keen interest in languages (including Sanskrit). André Weil soon specialized in mathematics, his sister in philosophy. He graduated (and was first in his year in the *agrégation* for mathematics) from the École Normale Supérieure (ENS) when he was not even nineteen years old, and traveled in Italy and Germany. He obtained his doctorate in Paris at the age of twenty-two, and then went to Aligarh, India, as a professor for two years. After a brief spell in Marseilles, he was *Maitre de Conférences* at Strasbourg University (along with Henri Cartan) from 1933 to 1939. The idea of the Bourbaki project arose there from discussions about teaching with Cartan, and grew in Paris in meetings that included other friends from the ENS.

His research achievements began with his 1928 Paris thesis. In it, he generalized MORDELL’S THEOREM [V.32] of 1922, that the group of rational points on an ELLIPTIC CURVE [III.21] is a finitely generated Abelian group, to the group of K -rational points (where K is a NUMBER

FIELD [III.65]) of a Jacobian variety. During the following twelve years, Weil branched out in various directions, all related to important research topics of the 1930s: the approximation of holomorphic functions of several variables by polynomials; the conjugation of maximal tori in compact LIE GROUPS [III.50 §1]; the theory of integration on compact and Abelian topological groups; and the definition of uniform TOPOLOGICAL SPACES [III.92]. But problems of arithmetic origin stood out among his interests: further thoughts on his thesis and on Siegel’s finiteness theorem for integral points; a bold “vector bundle” version of the RIEMANN–ROCH THEOREM [V.34] on a Riemann surface (in parallel with similar work by E. Witt); p -adic analogs of ELLIPTIC FUNCTIONS [V.34] (with his student Elisabeth Lutz).

Starting in 1940, Weil became active on what was probably the biggest challenge in arithmetic algebraic geometry at the time. Helmut Hasse had proved in 1932 the analogue of THE RIEMANN HYPOTHESIS [IV.2 §3] for curves of genus 1 (elliptic curves) defined over a field with finitely many elements. The problem was to generalize this to algebraic curves of genus higher than 1. In 1936, Max Deuring had proposed algebraic correspondences as a crucial new ingredient for attacking this problem; but the problem remained open until World War II. Weil’s initial attempt, written while in jail in Rouen, was very modest, and contained little more than Deuring’s observations of 1936. But, after several years of searching in various directions while in residence in the United States, Weil finally became the first person to prove the analogue of the Riemann hypothesis for all nonsingular curves. This proof relied on his complete rewriting of algebraic geometry (over an arbitrary ground field), which he had published before in his *Foundations of Algebraic Geometry* (1946). Furthermore, Weil generalized the analogue of the Riemann hypothesis from curves to algebraic varieties of arbitrary dimensions, defined over a finite field, and added a new topological interpretation of the main invariants of the relevant zeta functions. Taken together, all these conjectures became known as THE WEIL CONJECTURES [V.38]; they represented the most important stimulus for the further development of algebraic geometry right through to the 1970s, and to some extent later as well.

Several mathematicians were at work in the 1930s and 1940s trying to rewrite algebraic geometry. Weil’s *Foundations*, even though it does contain striking new insights (e.g., a novel definition of intersection multiplicity), owes its basic notions (generic points, specializations) to van der Waerden, and it exerted its

influence on the mathematical community in conjunction with the (different) rewriting of algebraic geometry developed so successfully by Oscar Zariski from 1938 onward. It was therefore to a large extent the characteristic style, rather than just the “mathematical content,” of the *Foundations* that would create a new way of doing algebraic geometry for the next twenty years or so, until it began to be replaced by Grothendieck’s language of schemes.

Later works include, among other seminal papers and books, Weil’s “adelic” rewriting of Siegel’s work on quadratic forms, and a crucial contribution to the philosophy, due to Taniyama and Shimura, that elliptic curves over the rational numbers should be modular—the proof of this fact is the basis of Wiles’s 1995 proof of FERMAT’S LAST THEOREM [V.12].

In 1947, Weil—whose evasion of the French draft in 1939 was considered very critically by many American colleagues—finally obtained a professorship at a distinguished university, namely Chicago. In 1958, he moved to Princeton as a permanent member of the Institute for Advanced Study.

The postwar years saw Weil continuously active on many fronts of mathematical research, contributing insightfully to many subjects that were in the air at the time. To mention just a few: the Weil groups of CLASS FIELD THEORY [V.31]; the explicit formulas of analytic number theory; various aspects of differential geometry, in particular KÄHLER MANIFOLDS [III.90 §3]; the determination of Dirichlet series by their functional equations. All of these topics point to seminal works without which today’s mathematics would not be what it is.

In his later years, Weil put his erudition and historical sense to work writing articles and a book on the history of mathematics: *Number Theory, an Approach through History*. He also published a partial autobiography ending in 1945, *Souvenirs d’Apprentissage*, of considerable literary quality.

Further Reading

- Weil, A. 1976. *Elliptic Functions According to Eisenstein and Kronecker*. Ergebnisse der Mathematik und ihrer Grenzgebiete, volume 88. Berlin: Springer.
- . 1980. *Oeuvres Scientifiques/Collected Papers*, second edn. Berlin: Springer.
- . 1984. *Number Theory. An Approach through History. From Hammurapi to Legendre*. Boston, MA: Birkhäuser.
- . 1991. *Souvenirs d’Apprentissage*. Basel: Birkhäuser 1991. (English translation: 1992, *The Apprenticeship of a Mathematician*. Basel: Birkhäuser.)

Norbert Schappacher and Birgit Petri

VI.94 Alan Turing

b. London, 1912; d. Wilmslow, England, 1954

Logic; computing; cryptography; mathematical biology

In 1936, as a young Fellow of King’s College, Cambridge, Alan Turing made a crucial contribution to mathematical logic: he defined “computability” with what is now called the TURING MACHINES [IV.20 §1.1]. Although mathematically equivalent to a definition of effective calculability earlier given by CHURCH [VI.89], Turing’s concept was compelling because of his entirely original philosophical analysis. It won the endorsement of Church, and indeed also of GÖDEL [VI.92], whose 1931 INCOMPLETENESS THEOREM [V.18] underlay Turing’s investigation. Using his definition, Turing showed that first-order logic was undecidable, and thus dealt the final death blow to HILBERT’S [VI.63] formalist program. (See LOGIC AND MODEL THEORY [IV.23 §2].)

Computability is now fundamental in mathematics, in that it gives an exact meaning to the question of whether a method exists to solve a problem. As an illustration, HILBERT’S TENTH PROBLEM [V.23], on the general solubility of Diophantine equations, was completely resolved in 1970 by methods connected with Turing’s ideas. Turing himself pioneered extensions of his definition in mathematical logic, and applications of it in algebra. However, he was unusual as a mathematician in that he explored not only the mathematical uses of his ideas (in questions of decidability in algebra) but also the wider implications for philosophy, science, and engineering.

One factor in Turing’s breakthrough was his fascination with the problem of mind and matter. Turing’s analysis of mental states and operations has since become a point of departure for the cognitive sciences. Turing himself blazed this trail later by his advocacy of the possibility of artificial intelligence. His famous 1950 “Turing test” was part of an extensive range of research proposals in this field.

A more immediately applicable aspect of his 1936 work lay in his observation that a single “universal” machine could do the work of any Turing machine, by reading the description of that machine as a table of instructions. This is the essential principle of the modern digital computer, whose programs are themselves data structures. In 1945 Turing used this insight to plan a first electronic computer and its programming. He

was preempted by VON NEUMANN [VI.91], but it can be argued that von Neumann had used Turing's insight that computing must be primarily an application of logic. Thus, Turing laid the foundations of modern computer science.

Turing was able to bridge theory and practice because between 1938 and 1945 he was the chief scientific figure in British cryptography, with particular responsibility for decrypting German naval signals. His main contributions lay in a brilliant logical solution of the Enigma cipher, and in Bayesian information theory. The advanced electronics employed in British code breaking gave him the experience to become a pioneer of practical computing as well.

Turing had less success in postwar computer engineering, and increasingly withdrew from attempts to influence the course of computer development. Instead, at Manchester University after 1949 he concentrated on a theory of nonlinear partial differential equations applied to biological development. Like his 1936 work, this opened an entirely new field. It also illustrated his broad mathematical scope, which included important work on the RIEMANN ZETA FUNCTION [IV.2 §3]. He was busy working on biological theory and new ideas in physics at the time of his sudden death.

Turing's short life combined the purest mathematics and the most practical applications. It was also marked by other contrasts. Although he promoted the theme of computer-based artificial intelligence, there was nothing mechanical about his thought or life. The wit and drama of the "Turing test" have made him a lasting figure in the popularization of mathematical ideas. The dramatization of his life, drawing on the extraordinary secrecy of his war work, and his subsequent persecution as a homosexual, have also attracted great public interest.

Further Reading

- Hodges, A. 1983. *Alan Turing: The Enigma*. New York: Simon & Schuster.
- Turing, A. M. 1992–2001. *The Collected Works of A. M. Turing*. Amsterdam: Elsevier.

Andrew Hodges

- b. Waldenburg (Lower Silesia; now Walbrzych, Poland), 1918;
d. New Haven, Connecticut, 1974
Applied mathematics; logic; model theory; nonstandard analysis

Robinson was educated at a private Rabbinical school and then at the Jewish High School in Breslau until 1933, when he emigrated with his family to Palestine. There Robinson finished high school, going on to study mathematics at the Hebrew University under Abraham Fraenkel. He spent the spring of 1940 at the Sorbonne, but when the Germans invaded France Robinson made his way to England. There he spent the war as a refugee, in the service of the Free French Forces. Robinson's mathematical talents were soon recognized, and he was assigned to the Royal Aircraft Establishment in Farnborough, where he was part of a team designing supersonic delta wings and reconstructing German V-2 rockets to determine how they worked. After the war, Robinson received his M.Sc. degree in mathematics from the Hebrew University, with minors in physics and philosophy. Several years later, he completed his Ph.D. in mathematics at Birkbeck College, London. His thesis, "On the metamathematics of algebra," was published in 1951.

Meanwhile, Robinson had been teaching at the Royal College of Aeronautics since its founding in Cranfield in October of 1946. Although promoted to Deputy Head of the Department of Aeronautics in 1950, in the following year Robinson accepted a position, at the rank of associate professor, at the University of Toronto in the Department of Applied Mathematics. While at Toronto, most of his publications were devoted to applied mathematics, including papers on supersonic airfoil design and a book he coauthored with his former student from Cranfield, J. A. Laurmann, on *Wing Theory*.

His years at Toronto (1951–57) proved to be a transitional period in Robinson's career, as his interests turned increasingly toward mathematical logic, beginning with studies of algebraically closed fields of characteristic zero. In 1955 he published a book in French summarizing much of his early work in mathematical logic and MODEL THEORY [IV.23], *Théorie Métamathématique des Ideaux*. Robinson was a pioneering contributor to model theory, which at its simplest uses mathematical logic to analyze mathematical structures (like groups, fields, or even set theory itself). Given an axiomatic system, a *model* is a structure that satisfies the axioms. One of his early impressive results was a model-theoretic proof, which he published in 1955 in *Mathematische Annalen*, of Hilbert's seventeenth problem, namely that a positive-definite rational

function over the reals can be expressed as a sum of squares of rational functions. This was soon followed by another book, *Complete Theories* (1956), which further extended ideas he had explored earlier in his thesis on model-theoretic algebra. Here Robinson introduced such important concepts as model completeness, model completion, and the “prime model test,” along with proofs of the completeness of REAL-CLOSED FIELDS [IV.23 §5] and the uniqueness of the model completion of a model-complete theory.

In the fall of 1957 Robinson returned to the Hebrew University, where he assumed the chair formerly held by his teacher Abraham Fraenkel in the Einstein Institute of Mathematics. While at the Hebrew University, Robinson worked on aspects of local differential algebra, differentially closed fields, and in logic on SKOLEM’s [VI.81] results dealing with nonstandard models of arithmetic. These provide models of ordinary PEANO ARITHMETIC [III.69], the usual arithmetic of the integers $(0, 1, 2, 3, \dots)$, but ones that include “nonstandard” elements, “numbers” that extend the scope of the standard model to models that are larger but nevertheless satisfy the axioms of the standard structure. A nonstandard model of arithmetic may include, for example, infinite integers. As Haim Gaifman puts it succinctly, “A nonstandard model is one that constitutes an interpretation of a formal system that is admittedly different from the intended one.”

Robinson spent the year 1960–61 in the United States, at Princeton, replacing CHURCH [VI.89], who was on sabbatical leave. It was there that Robinson was inspired to make his most revolutionary contribution to mathematics, nonstandard analysis, using model theory to allow the rigorous introduction of infinitesimals. In fact, this extended the usual, standard model of the real numbers to a nonstandard model that included both infinite and infinitesimal elements. He first published on this topic in 1961 in the *Proceedings of the Netherlands Royal Academy of Sciences*. This paper was soon followed by a book, *Introduction to Model Theory and to the Metamathematics of Algebra* (1963), a thorough revision of his earlier book of 1951, including a new section on nonstandard analysis.

Meanwhile, Robinson had left Jerusalem for Los Angeles, where he was appointed as Carnap’s chair at UCLA in mathematics and philosophy. In addition to writing an introductory text, *Numbers and Ideals: An Introduction to Some Basic Concepts of Algebra and Number Theory* (1965), he also published his definitive introduction to *Nonstandard Analysis* (1966). Among

the important results he obtained while at UCLA (1962–67) was his proof of the invariant subspace theorem in Hilbert space for the case of polynomially compact operators, published with his graduate student Allen Bernstein. (The case for compact operators had been established by Aronszajn and Smith in 1954; what Bernstein and Robinson did was extend this to the case of an operator T such that some nonzero polynomial of T is compact.)

In 1967 Robinson moved to Yale University (1967–74), where he was eventually given a Sterling Professorship in 1971. Among Robinson’s most important mathematical achievements during this period were his extension of Paul Cohen’s method of FORCING [IV.22 §5.2] in set theory to model theory, and applications of nonstandard analysis in economics and quantum physics. He also applied nonstandard analysis to achieve an outstanding result in number theory, namely a simplification of Carl Ludwig Siegel’s theorem regarding integer points on curves (1929), as generalized by Kurt Mahler for rational as well as integer solutions (1934). This was work that Robinson did jointly with Peter Roquette; together they extended the Siegel–Mahler theorem by considering nonstandard integer points and nonstandard prime divisors. After Robinson’s death from pancreatic cancer in 1974, Roquette published this work in the *Journal of Number Theory* in 1975.

Further Reading

- Dauben, J. W. 1995. *Abraham Robinson. The Creation of Nonstandard Analysis. A Personal and Mathematical Odyssey*. Princeton, NJ: Princeton University Press.
- . 2002. Abraham Robinson. 1918–1974. *Biographical Memoirs of the National Academy of Sciences* 82:1–44.
- Davis, M., and R. Hersh. 1972. Nonstandard analysis. *Scientific American* 226:78–86.
- Gaifman, H. 2003. *Non-Standard Models in a Broader Perspective. Nonstandard Models of Arithmetic and Set Theory*, edited by A. Enayat and R. Kossak, pp. 1–22. Providence, RI: American Mathematical Society.

Joseph W. Dauben

VI.96 Nicolas Bourbaki

b. Paris, 1935; d. —

Set theory; algebra; topology; foundations of mathematics; analysis; differential and algebraic geometry; integration theory; spectral theory; Lie algebras; commutative algebras; history of mathematics

Anne: you suggested ‘fl.’ instead of ‘b.’ here, but the problem is that the group is still ‘flourishing’ so that won’t work. OK to keep it as it is?

Bourbaki is a pseudonym chosen in 1935 by a group of French mathematicians, including Henri Cartan, Jean Dieudonné, and ANDRÉ WEIL [VI.93]. Under this nom de plume, several generations of mostly French mathematicians conceived, wrote, and published a series of treatises under the general title *Éléments de Mathématique*. The uncommon use of the singular “mathématique” underscored a strong commitment to the unity of mathematics that is one of the chief characteristics of the group. Together with the “Bourbaki Seminar,” this monumental work promoted a unified, axiomatic, and structural view of pure mathematics that has exerted a strong influence on teaching and research since World War II, especially in France.

Charles Denis Sauter Bourbaki was a French general who fought in the Franco-Prussian war in 1870–71. A hoax lecture given by students at the École Normale Supérieure to the entering class in 1923 culminated with a “Bourbaki theorem.” In 1935, a group of mathematicians, many of whom had taken part in that lecture, as either audience or pranksters, decided to adopt that name for the fictive author of the modern treatise of analysis they were planning to write.

Their first meeting had taken place in Paris on December 10, 1934. In addition to Cartan, Dieudonné, and Weil, other young university professors of mathematics were present: Claude Chevalley, Jean Delsarte, and René de Possel. Agreeing that analysis textbooks available in French (such as Édouard Goursat’s *Cours d’Analyse*) were outdated, they decided to write a book, collectively, to replace them. Having been in touch with modern German mathematics, especially at HILBERT’S [VI.63] Göttingen, and influenced in particular by Bartel van der Waerden’s *Moderne Algebra*, they thought that their large treatise should begin with an “abstract packet” summarizing in axiomatic form basic general notions such as sets, groups, and fields. Soon after this, Szolem Mandelbrojt joined the group. Paul Dubreil and Jean Leray took part in just a few of the original meetings, and were replaced by Charles Ehresmann and the physicist Jean Coulomb.

In July 1935, the group had its first “congress” (as its annual summer meetings would later be called) in Besse-en-Chandesse, Auvergne, where the pen name “N. Bourbaki” was definitively adopted (the first name, Nicolas, was chosen later). Settling on working procedures, they drew up the general outline of the planned treatise. The members of the group worked collectively following certain ritual rules. They co-opted new collaborators, kept membership secret, and refused to

acknowledge individual contributions. During the three or four working sessions they held every year, each contribution prepared in advance by one of them was read line by line, discussed, and severely criticized by the others. Up to ten successive drafts and several years of work by various authors were often needed before a final version was unanimously adopted.

The first booklet—a digest of results in set theory—was dated 1939 but issued in 1940. Despite the difficult working conditions during World War II, this was soon followed in the 1940s by several booklets dealing mostly with general topology and algebra. Today, the *Elements of Mathematics* consists of several books: *Theory of Sets, Algebra, General Topology, Real-Variable Functions, Topological Vector Spaces, Integration, Commutative Algebra, Differential and Analytic Manifolds, Lie Groups and Lie Algebra, Spectral Theories*, and *Elements of the History of Mathematics*. Many of them have been extensively revised over the years and translated into several languages, including English and Russian.

The first six books formed a tight linear exposition entitled “The fundamental structures of analysis.” When they first appeared, they were striking for the logical organization of the topics covered. The axiomatic method was used systematically, and great effort was made to ensure a global unity of style, notation, and terminology. The avowed ambition was to take mathematics from its very start and, proceeding from the general toward the particular, write a unified survey of most of modern mathematics.

Several generations of mathematicians were co-opted into the “Association of Bourbaki’s Collaborators,” as the group is now officially known. After World War II, Samuel Eilenberg, Laurent Schwartz, Roger Godement, Jean-Louis Koszul, and Jean-Pierre Serre, among others, took part in the writing of the treatise. Later, Armand Borel, John Tate, François Bruhat, Serge Lang, and Alexander Grothendieck also joined. Although its frequency of publication has now slowed to a trickle, the group is still functioning in the first decade of the twenty-first century.

Notwithstanding the number of collaborators involved and the extensiveness of the work published, Bourbaki’s vision of mathematics was, and has remained, surprisingly coherent. Most of the crucial mathematical choices, which would come to have a huge impact on the structural image of mathematics that the group would later vigorously promote, were made in the late 1930s. In the follow-

ing decades, many mathematicians shared a conviction that a tight axiomatic refoundation of their research domains would help overcome current blockages. This was felt, for example, in probability theory, model theory, algebraic geometry and topology, commutative algebra, Lie groups, and Lie algebras.

After World War II, as the notoriety of both the group and its individual members steadily grew, Bourbaki's public image soon encompassed more than just the treatise. At the level of mathematical research, the Bourbaki Seminar was a prestigious outlet established in Paris in 1948, and it has met three times a year ever since. Members of Bourbaki selected speakers who usually summarized someone else's work, and supervised the publication of their talks. The topics selected emphasized specific domains of mathematics, such as algebraic and differential geometry, at the expense of others, such as probability theory or applied mathematics.

Bourbaki's views on the philosophy of mathematics were always clear, especially after two articles published in the late 1940s under that name argued for a complete reorganization of mathematics, eschewing older classification schemes in favor of fundamental structures (sometimes called "mother-structures" and supposedly closer to the deep mental structures of humans) meant to underscore the organic unity of mathematics. Bourbaki's public image was echoed by structuralists in the human sciences as well as artists and philosophers, and it was invoked by radical reformers of mathematical education from kindergarten to university—although actual members of Bourbaki were rarely involved directly.

From the late 1960s, Bourbaki's critics became louder on two counts: they took issue with the Bourbaki approach to the logical foundations of mathematics and they found gaps in the group's encyclopedic objectives. CATEGORY THEORY [III.8] developed by Saunders Mac Lane and Samuel Eilenberg was found to offer a more fruitful foundational framework than Bourbaki's structures. It also became clear that whole branches of mathematics—probability theory, geometry, and, to a lesser extent, analysis and logic—were to remain absent from the treatise, their very place in the grand architecture of Bourbakist mathematics left unclear. For a new generation of mathematicians, it was Bourbaki's elitist contempt for applications that was especially damaging.

Bourbaki's impact on mathematics was profound: despite its excesses, Bourbaki's unified, structural, rig-

orous image of mathematics is still with us. But it was those very characteristics that led to a feeling that Bourbaki was corseting mathematical research. The backlash seems to be abating somewhat nowadays, but no new Bourbaki is in view.

Further Reading

- Beaulieu, L. 1994. Questions and answers about Bourbaki's early work (1934–1944). In *The Intersection of History and Mathematics*, edited by S. Chikara et al., pp. 241–52. Basel: Birkhäuser.
- Corry, L. 1996. *Modern Algebra and the Rise of Mathematical Structures*. Basel: Birkhäuser.
- Mac Lane, S. 1996. Structures in mathematics. *Philosophia Mathematica* 4:174–86.

David Aubin

Part VII

The Influence of Mathematics

VII.1 Mathematics and Chemistry

Jacek Klinowski and Alan L. Mackay

1 Introduction

Since ARCHIMEDES [VI.3], and his experimental investigation (described by Vitruvius) of the proportions of gold and silver in an alloy, the solution of chemical problems has employed mathematics. Carl Schorlemmer studied the paraffinic series of hydrocarbons (then important because of the discovery of oil in Pennsylvania) and showed how their properties changed with the addition of successive carbon atoms. His close friend in Manchester, Friedrich Engels, was inspired by this to introduce the transformation of “quantity into quality” into his philosophical outlook, which then became a mantra of dialectical materialism. From a similar chemical observation, CAYLEY [VI.46] in 1857 developed “rooted trees” and the mathematics of the enumeration of branched molecules, the first articulation of GRAPH THEORY [III.34]. Later, George Pólya developed his fundamental enumeration theorem, facilitating further advances in the counting of these molecules. Still more recently, chemical problems such as the mechanics and kinematics of DNA have had a significant influence on KNOT THEORY [III.46].

However, chemistry has been a quantitative modern science for no more than 150 years. Before this, it was a distant dream: when NEWTON [VI.14] was developing the calculus in around 1700, much of his time was spent working on alchemy. He explained why, having established “the motions of the planets, the comets, the Moon and the sea,” he was unable to determine the remaining structure of the world from the same propositions:

I suspect that they may all depend upon certain forces by which the particles of the bodies, by some causes hitherto unknown, are either mutually impelled toward

one another, and cohere in regular figures, or are repelled and recede from one another. These forces being unknown, philosophers have hitherto attempted the search of Nature in vain; but I hope the principles laid down will afford some light either to this or some truer method of philosophy.

The nature of such forces came to be understood only two hundred years later, and indeed the electron, the particle responsible for chemical bonding, was not discovered until 1897. This is why the main flow of ideas has been from mathematical theory to applications in chemistry.

Some of the fundamental equations of chemistry, though based on experiment rather than strict mathematical reasoning, convey a wealth of information with great simplicity and elegance (Thomas 2003). For example, consider Boltzmann’s fundamental equation of statistical thermodynamics, which links entropy, S , to Ω , the number of possible ways of arranging the particles: $S = k \log \Omega$, where k is known as the Boltzmann constant. There is also the expression derived by Balmer for the wavelength, λ , of spectral lines from hydrogen in the visible portion of the spectrum:

$$\frac{1}{\lambda} = R \left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right),$$

where n_1 and n_2 are integers, $n_1 < n_2$, and R is known as the Rydberg constant. A third example, the Bragg equation, links the wavelength, λ , of monochromatic X-rays, the distance, d , between planes in a crystal lattice, and the angle, θ , between the crystal planes and the direction of the X-rays. It says that $n\lambda = 2d \sin \theta$, where n is a small integer. Finally, there is the “phase rule,” $P + F = C + 2$, which links the number of phases, P , the number of degrees of freedom, F , and the number of components, C , in a chemical system. This is the same relationship as that between the number of vertices, faces, and edges in a convex polyhedron, and emerges from the geometrical representation of the system.

In recent years computers have become the dominant tool in theoretical chemistry. Not only can computers solve differential equations numerically, they can often provide exact algebraic expressions, sometimes even ones that are too elaborate to write out. Computing has required the development of algorithms in the fields of *structure, process, modeling, and search*. Mathematics has been revolutionized by the advent of computers: in particular in the facility for dealing with nonlinear problems and for displaying results graphically. This has led to fundamental advances, some of them bearing on chemistry.

In general, mathematical approaches to chemical problems can be divided into discrete and continuous treatments, reflecting on the one hand the fundamental discrete atomic nature of matter and on the other the continuous statistical behavior of large numbers of atoms. For example, enumerating molecules is a discrete problem, while a problem involving global measures such as temperature and other thermodynamic parameters will be continuous. These treatments have required different branches of mathematics, with integers more important for discrete problems and real numbers more important for continuous ones.

We shall now outline some chemical problems to which, in our view, mathematics has made the most significant contributions.

2 Structure

2.1 Description of Crystal Structure

Crystal structure is the study of how atoms arrange themselves to form macroscopic materials. Early ideas in the subject were based purely on the symmetry of crystals and their morphology (that is, the shapes they tended to form), and were developed in the nineteenth century in the absence of definite information about the atomic structure of matter. The 230 *space groups*, which codify different ways of arranging objects periodically in three-dimensional (3D) space, were found independently by Fedorov, Schoenflies, and Barlow between 1885 and 1891. They result from the systematic combination of a certain collection of fourteen lattices, named *Bravais lattices* after their discovery in 1848 by Auguste Bravais, with the thirty-two so-called *crystallographic point groups*, which were developed from morphological considerations.

Since the diffraction of X-rays was demonstrated in 1912 by Max von Laue and practical X-ray analysis was developed by W. H. Bragg and his son W. L. Bragg, the

crystal structures of several hundred thousand inorganic and organic substances have been determined. However, such analysis was for a long time held back by the time required for the calculation of FOURIER TRANSFORMS [III.27]. This difficulty is now a thing of the past, owing to the discovery of THE FAST FOURIER TRANSFORM [III.26] by Cooley and Tukey in 1965—a universally applied algorithm and one of those most often cited in mathematics and computer science.

The fundamental geometry of two-dimensional (2D) and 3D spatial structures led mathematicians to seek analogous problems in N dimensions. Some of this work has found application in the description of *quasicrystals*, which are arrangements of atoms that, like crystals, exhibit a high degree of organization, but which lack the periodic behavior of crystals. (That is, they do not have translational symmetry.) The most notable example is the following, which uses six-dimensional geometry. Take a regular cubic lattice L in six dimensions and let V be a 3D subspace of \mathbb{R}^6 that contains no point of L apart from the origin. Now project on to V all points from L that are closer to V than a certain distance d . The result is a 3D structure of points that exhibits a great deal of local regularity but not global regularity. This structure gives a very good model for quasicrystals.

Until recently, crystals in three dimensions had always been thought of as periodic, and therefore capable of showing only twofold, threefold, fourfold, or sixfold axes of symmetry. Fivefold axes were excluded, because a plane cannot be tiled with regular pentagons. However, in 1982, X-ray and electron diffraction demonstrated the presence of fivefold diffraction symmetry in certain rapidly cooled alloys. Careful electron microscopy was necessary to distinguish the observed structures from the twinning (symmetrical intergrowth) of “normal” crystals. This discovery, of a quasicrystalline alloy phase “with long-range orientational order and no translational symmetry,” has brought about an ideological shift in crystallography.

The earlier concept of a “quasilattice” appeared to be one possible mathematical formalism for the description of quasicrystals. Quasilattices have two incommensurable periods in the same direction, and the ratio of these periods was given by so-called Pisot and Salem numbers. A *Pisot number* θ is a root of a polynomial with integer coefficients of degree m such that if $\theta_2, \dots, \theta_m$ are the other roots, then $|\theta_i| < 1$, $i = 2, \dots, m$. A real quadratic algebraic integer (see ALGEBRAIC NUMBERS [IV.1 §11]) greater than 1 and of

degree 2 or 3 is a Pisot number if its norm is equal to ± 1 . The golden ratio is an example of a Pisot number since it has degree 2 and norm -1 . A *Salem number* is defined in a similar way to a Pisot number, but with the inequalities replaced by equalities.

LIE ALGEBRA [III.50 §2] arguments have also been used to describe quasicrystals. This has stimulated a great deal of theoretical N -dimensional geometry. Before the discovery of quasicrystals, Roger Penrose had shown how to cover a plane nonperiodically using two different types of rhombic tiles, and corresponding rules were developed for 3D space with two kinds of rhombohedral tiles. The Fourier transform of such a 3D structure with atoms placed in the rhombohedral cells explains the observed diffraction patterns of 3D quasicrystals, while Penrose's 2D pattern corresponds to *decagonal quasicrystals*, which consist of stacked layers of the 2D pattern and which have been experimentally observed.

The broadening of classical crystallography to encompass quasicrystals has been given further impetus by recent advances in electron microscopy. It is now possible to observe atomic arrangements directly, including those of the decagonal quasicrystals just mentioned, rather than having to deduce them from diffraction patterns, where the phases of the various diffracted beams are lost in the experimental system and have to be recovered mathematically. The whole field of computational and experimental image processing has become coherent as a result.

Another model describes 2D quasicrystals in terms of a single repeating unit, but the unit is a composite object, a pattern made out of identical decagons. Unlike the unit cells in periodic crystals, these quasi-unit cells are allowed to overlap, but where they do their constituent decagons must match up. This conceptual device is an alternative to the use of two kinds of unit cell. It emphasizes the dominating physical presence of locally ordered atomic clusters, with no long-range order, and it can be extended to three dimensions. The predictions of this model agree with the observed composition of a 2D decagonal quasicrystal, as well as with the results obtained by electron microscopy and X-ray diffraction. Nevertheless, although a huge amount of interesting mathematics has been generated by the discovery of quasicrystals, most of it is not physically relevant: the structures emerge from the competition between local and global ordering forces rather than from the mathematics of the Penrose tiling.

The acceptance of quasicrystals demonstrates the need to accommodate more general concepts of *order* into classical crystallography. It has explicitly introduced concepts of *hierarchy*, by involving not just ordered clusters of atoms but ordered clusters of clusters, where local order has predominated over the regular lattice repetition. Quasicrystals represent the first step from absolute regularity toward more general structures that are intimately bound up with the notion of *information*.

Information can be stored in a device which has two or more clearly identifiable states that are *metastable*. This means that each state is a local equilibrium, and to pass from one to another, one must supply and remove enough energy to take the device over the local energy watershed. A switch, for example, can be on or off; it is stable in either state and to change the state takes a certain amount of energy. To take a more general example, any information, encoded as a sequence of binary digits, can be read in, read out, and stored as a sequence of magnetic domains, where each one is magnetized either north or south.

Perfect crystals have no alternative metastable states, so cannot be used to store information, but a piece of silicon carbide, for example, exists as a sequence of close-packed layers, each of which may be in one or other of two almost equivalent positions. To describe the structure of a piece of silicon carbide therefore demands a knowledge of the sequence of positions in which the layers are stacked. This can be represented by a string of binary digits. Now that it is possible to arrange atoms in a structure almost at will, at least if they are on a surface, the processing of information has become important to chemistry.

In determining the arrangement of atoms in crystals, mathematics has been essential for the solution of the *phase problem*, which had held up progress in structural chemistry and molecular biology for decades. A pattern of diffracted X-rays, recorded as an array of spots on a photographic plate, depends on the arrangement of atoms in the molecule causing the diffraction. The problem is that the diffraction pattern registers only the intensity of the light waves, but to work back to the molecular structure it is necessary to know their phase as well (that is, the positions of the crests and troughs of the waves relative to each other). This results in a classic *inverse problem*, which was solved by Jerome and Isabella Karle and Herbert A. Hauptman.

A *Voronoi diagram* consists of points, representing atom sites, with each point contained in a region (see

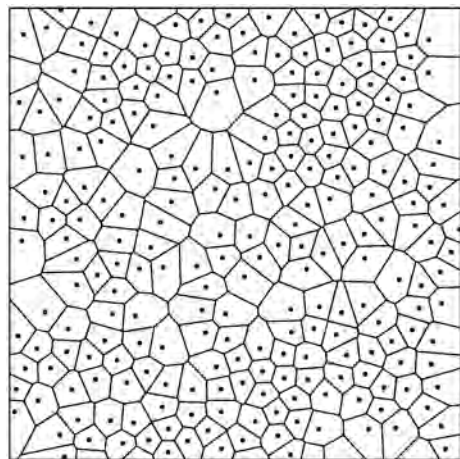


Figure 1 Voronoi dissection of 2D space.

also MATHEMATICAL BIOLOGY [VII.2 §5]). The region surrounding a given site consists of all points that are closer to that site than to any of the other sites (figure 1). The geometric dual of the Voronoi diagram, a system of triangles with the sites as vertices, is called the *Delaunay triangulation*. (An alternative definition of the Delaunay triangulation is that it is a triangulation of the sites with the additional property that, for each triangle, the circumcircle of that triangle contains no other sites.) These dissections give a well-defined way of representing many N -dimensional chemical structures as arrangements of polytopes. Crystals, which have periodic boundaries, are easier to deal with than extended structures that terminate in a boundary. The Voronoi dissection of crystal structures enables one to describe them as networks. Nevertheless, despite much progress in understanding structure, it is not yet possible to guess a crystal structure in advance just from the composition of elements in its molecules.

2.2 Computational Chemistry

Attempts to solve THE SCHRÖDINGER EQUATION [III.85], which gives the quantum mechanical description of matter, began soon after it was proposed in 1926. For very simple systems, calculations performed on mechanical calculators agreed with the experimental results of spectroscopy. In the 1950s, electronic computers became available for general scientific use, and the new field of *computational chemistry* developed, the aim of which was to obtain quantitative information on atomic positions, bond lengths, electronic configurations of atoms, etc., by means of numerical solu-

tions of the Schrödinger equation. Advances during the 1960s included deriving suitable functions for representing electronic orbitals, obtaining approximate solutions to the problem of how the motions of different electrons correlate with each other, and providing formulas for the derivative of the energy of a molecule with respect to the positions of the atomic nuclei. Powerful software packages became available in the early 1970s. Much current research is aimed at developing methods that can handle larger and larger molecules.

Density functional theory (DFT) (Parr and Yang 1989) is a major recent field of activity in quantum mechanical computation, and concerns macroscopic features of materials. It has been successful in the description of the properties of metals, semiconductors, and insulators, and even of complex materials such as proteins and carbon nanotubes. Traditional methods in the study of electronic structure—such as one called the *Hartree-Fock theory molecular orbital method*, which assigns the electrons two at a time to a set of molecular orbitals—involve very complicated many-electron wave functions. The main objective of DFT is to replace the many-body electronic wave function, which depends on $3N$ variables, with a different basic quantity, the *electronic density*, which depends on just 3 variables, and therefore greatly speeds up calculations.

The partial differential equations of quantum mechanics, physics, fields, surfaces, potentials, and waves can sometimes be solved analytically, but even if they cannot, they are now almost always soluble by numerical methods. All this relies on the corresponding pure mathematics (see NUMERICAL ANALYSIS [IV.21 §5]).

2.3 Chemical Topology

Isomers are chemical compounds that are made out of the same elements but have different physical and chemical properties. This can happen for various reasons. In *structural isomers*, the atoms and functional groups are linked together in different ways. This class includes *chain isomers*, where hydrocarbon chains have variable amounts of branching, and *position isomers*, where the position of a functional group in a chain is different (figure 2(a)). In *stereoisomers* the bond structure is the same, but the geometrical positioning of atoms and functional groups in space differs (figure 2(b)). This class includes *optical isomers*, where different isomers are mirror images of each other (figure 2(c)). While structural isomers have different

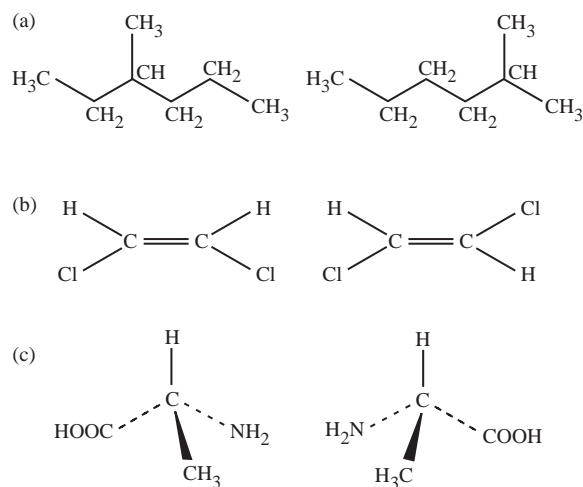


Figure 2 (a) Position isomerism. (b) Stereoisomerism. (c) Optical isomerism.

chemical properties, stereoisomers behave identically in most chemical reactions. There are also *topological isomers* such as catenanes and DNA.

An important theme in chemical topology is determining how many isomers there are of a given molecule. To do this, one first associates with any molecule a *molecular graph*, the vertices representing atoms and the edges representing chemical bonds. To enumerate *stereoisomers*, one counts the symmetries of this graph, but first one must consider symmetries of the molecule (Cotton 1990) in order to decide which symmetries of the graph correspond to spatial transformations that make chemical sense. Cayley addressed the problem of enumerating *structural isomers*, that is, combinatorially possible branched molecules. To do this, one must count how many different molecular graphs there are with a given set of elements, where two graphs are regarded as the same if they are isomorphic. The enumeration of isomorphism types uses group theory to count the intrinsic graph symmetries. After Pólya published his remarkable **ENUMERATION THEOREM** [IV.18 §6] in 1937, his work using **GENERATING FUNCTIONS** [IV.18 §§2.4, 3] and **PERMUTATION GROUPS** [III.70] became central to the enumeration of isomers in organic chemistry. The theorem solves the general problem of how many configurations there are with certain properties. It has applications such as the enumeration of chemical compounds and the enumeration of rooted trees in graph theory. A new branch of graph theory, called enumerative graph

theory, is based on Pólya's ideas (see **ALGEBRAIC AND ENUMERATIVE COMBINATORICS** [IV.18]).

Although not all the possible isomers occur in nature, molecules with remarkable topologies have been synthesized artificially. Among them are *cubane*, C_8H_8 , which contains eight carbon atoms arranged at the corners of a cube, each linked to a single hydrogen atom; *dodecahedrane*, $C_{20}H_{20}$, which, as its name suggests, has a dodecahedral shape; the *molecular trefoil knot*; and the self-assembling compound *olympiadane* composed of five interlocked rings. *Catenanes* (from Latin *catena*, chain) are molecules containing two or more interlocked rings that are inseparable without breaking a covalent bond. *Rotaxanes* (from Latin *rota*, wheel, and *axis*, axle) are dumbbell shaped, having a rod and two bulky stopper groups, around which there are encircling macrocyclic components. The stoppers of the dumbbell prevent the macrocycles from slipping off the rod. Even a molecular **MÖBIUS STRIP** [IV.7 §2.3] has recently been synthesized.

Macromolecules, such as synthetic polymers and biopolymers (e.g., DNA and proteins), are very large and highly flexible. The degree to which a polymer molecule coils and knots and links with other molecules is crucial to its physical and chemical properties, such as reactivity, viscosity, and crystallization behavior. The topological entanglement of short chains can be modeled using Monte Carlo simulation, and the results can now be experimentally verified with fluorescence microscopy.

DNA, the central substance of life, has a complex and fascinating topology, which is closely related to its biological function. The major geometric descriptions of supercoiled DNA (that is, DNA wrapped around a series of proteins) involve the concepts of linking, twisting, and writhing numbers that come from knot theory. DNA knots, which are created spontaneously within cells, interfere with replication, reduce transcription, and may decrease the stability of the DNA. "Resolvase enzymes" detect and remove these knots, but the mechanism of this process is not understood. However, using topological concepts of knots and tangles, one can gain insight into the reaction site and thereby try to infer the mechanism. (See also **MATHEMATICAL BIOLOGY** [VII.2 §5].)

2.4 Fullerenes

Graphite and diamond, the two crystalline forms of the element carbon, have been known since time immemorial, but *fullerenes*, which were subsequently found to

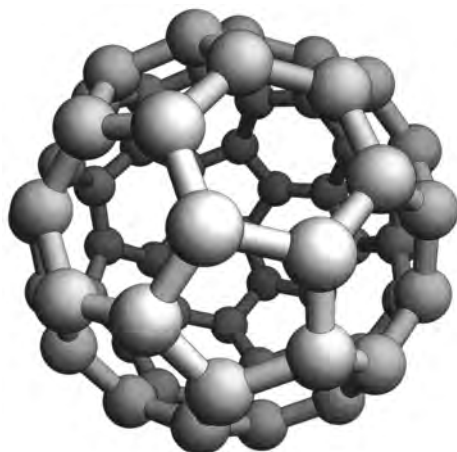


Figure 3 The structure of the fullerene C_{60} .

exist naturally in soot and geological deposits, were discovered only in the mid 1980s. The most common is the almost-spherical carbon cage C_{60} molecule (figure 3), also known as “buckminsterfullerene” after the architect who designed enormous domes, but fullerenes C_{24} , C_{28} , C_{32} , C_{36} , C_{50} , C_{70} , C_{76} , C_{84} , etc., also exist. Topology provides insights into the possible types of such structures, while group theory and graph theory describe the symmetry of the molecules, allowing one to interpret their vibrational modes.

In all fullerenes, each carbon atom is connected to exactly three neighboring ones, and the resulting molecule is a “cage” made of rings of either five or six carbon atoms. From EULER’s [VI.19] topological relationship $\sum_n (6 - n)f_n = 12$, where f_n is the number of n -hedral faces and the summation is over all faces of the polyhedron, we conclude first that $f_5 = 12$, since n is found to take only the values 5 or 6, and second that f_6 can take any value.

In 1994, Terrones and Mackay predicted the existence of ordered structures of a new kind, derived from graphite and related to fullerenes, with topologies of triply periodic MINIMAL SURFACES [III.96 §3.1]. These new structures, which are of great practical interest, are produced by introducing eight-membered rings of carbon atoms into a sheet of six-membered rings. This gives rise to saddle-shaped surfaces of negative GAUSSIAN CURVATURE [III.80], unlike the fullerenes, which have positive curvature. Thus, to model them mathematically one must consider embeddings of non-Euclidean 2D spaces into \mathbb{R}^3 . This has contributed to

a renewed interest in certain aspects of non-Euclidean geometry.

2.5 Spectroscopy

Spectroscopy is the study of the interaction of electromagnetic radiation (light, radio waves, X-rays, etc.) with matter. The central portion of the electromagnetic spectrum—spanning the infrared, visible, and ultraviolet wavelengths and the radio frequency region—is of particular interest to chemistry. A molecule, which consists of electrically charged nuclei and electrons, may interact with the oscillating electric and magnetic fields of light and absorb enough energy to be promoted from one discrete vibrational energy level to another. Such a transition is registered in the infrared spectrum of the molecule. The *Raman spectrum* monitors inelastic scattering of light by molecules (that is, when some of the light is scattered at a different frequency from the frequency of the incoming photons). Visible and ultraviolet light can redistribute the electrons in the molecule: this is *electronic spectroscopy*.

Group theory is essential in the interpretation of the spectra of chemical compounds (Cotton 1990; Hollas 2003). For any given molecule, the symmetry operations that can be applied to it form a GROUP [I.3 §2.1], and can be represented by matrices. This allows one to identify “spectroscopically active” events in a molecule. For example, just three bands are observed in the infrared spectrum and eight bands in the Raman spectrum of dodecahedrane. This is a consequence of the icosahedral symmetry of the molecule and is what one expects from group-theoretic considerations. Also, there are no coincidences between the infrared- and Raman-active modes. Similarly, group theory correctly predicts that, because of the high symmetry of a C_{60} molecule, it has only four lines in its infrared spectrum and ten in its Raman spectrum, even though it has 174 vibrational modes.

2.6 Curved Surfaces

Structural chemistry has greatly changed in the last twenty years. First, as we have seen, the rigid concept of a “perfect crystal” has been relaxed to embrace structures such as quasicrystals and textures. Second, an advance has been made from classical geometry to 3D differential geometry. The main reason for this has been the use of curved surfaces for describing a great variety of structures (Hyde et al. 1997).

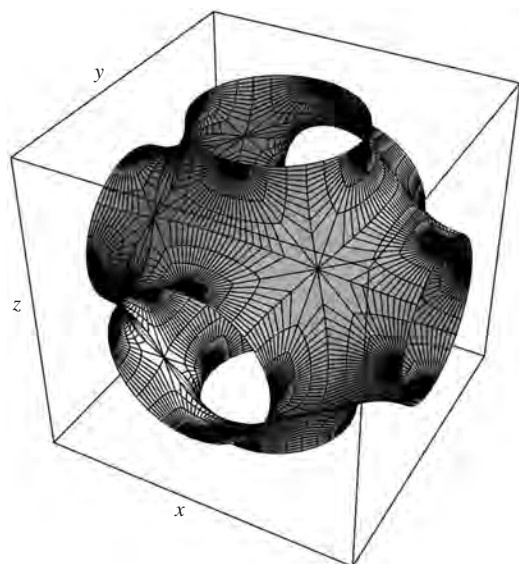


Figure 4 One unit cell of the P triply periodic minimal surface. The surface divides space into two interpenetrating labyrinths.

When a wire frame is dipped into soapy water, a thin film is formed. Surface tension minimizes the energy of the film, which is proportional to its surface area. As a result, the film has the smallest area consistent with the shape of the frame and with the requirement that the *mean curvature* of the film be zero at every point. If the symmetries of a minimal surface are given by one of the 230 space groups mentioned earlier, then the surface is periodic in three independent directions. Such triply periodic minimal surfaces (TPMSs) are of special interest because they appear in a variety of real structures such as silicates, bicontinuous mixtures, lyotropic colloids, detergent films, lipid bilayers, polymer interfaces, and biological formations (an example of a TPMS is illustrated in figure 4). Thus, TPMSs provide a concise description of many seemingly unrelated structures. Extensions of TPMSs may even have applications in cosmology as “branes.”

In 1866 WEIERSTRASS [VI.44] discovered a method of complex analysis suitable for general investigation of minimal surfaces. Consider a transformation of a minimal surface into the complex plane by combination of two simple maps. The first is the *Gauss map* ν , under which the image of a point P of the surface is the point P' of the intersection of the surface normal vector at P with the unit sphere centered at P . The second map is a stereographic projection σ of the point P' on the sphere

into the complex plane \mathbb{C} , resulting in the point P'' . The composite map, $\sigma\nu$, conformally maps the neighborhood of any nonumbilic point on the surface to a simply connected region of \mathbb{C} . (An umbilic point is one where the two principal curvatures are the same.) The inverse of this composite map is called the *Enneper-Weierstrass representation*.

In a system with the origin at (x_0, y_0, z_0) , the Cartesian coordinates (x, y, z) of *any* nontrivial minimal surface are determined by a set of three integrals:

$$\begin{aligned} x &= x_0 + \operatorname{Re} \int_{\omega_0}^{\omega} (1 - \tau^2) R(\tau) d\tau, \\ y &= y_0 + \operatorname{Re} \int_{\omega_0}^{\omega} i(1 + \tau^2) R(\tau) d\tau, \\ z &= z_0 + \operatorname{Re} \int_{\omega_0}^{\omega} 2\tau R(\tau) d\tau. \end{aligned}$$

Here $R(\tau)$ is the *Weierstrass function*. It is a function of a complex variable τ , and it is HOLOMORPHIC [I.3 §5.6] in a simply connected region of \mathbb{C} , except at isolated points.

The Cartesian coordinates of any (nonumbilic) point on a minimal surface are thus expressed as the real parts of certain contour integrals, evaluated in the complex plane from some fixed point ω_0 to a variable point ω . Integration is carried out within the domain where the integrands are analytic, and thus by CAUCHY'S THEOREM [V.6] the values of the integrals are independent of the path of integration from ω_0 to ω . In this way, a specific minimal surface is completely defined by its Weierstrass function.

While the Weierstrass functions for many TPMSs are unknown, the coordinates of points lying on *some* minimal surfaces involve functions of the form

$$R(\tau) = \frac{1}{\sqrt{\tau^8 + 2\mu\tau^6 + \lambda\tau^4 + 2\mu\tau^2 + 1}},$$

where μ and λ are sufficient to parametrize the surface. A method has been developed for deriving this function for a given type of surface, and it generates different families of minimal surfaces from the above equation. For example, taking $\mu = 0$ and $\lambda = -14$ gives a surface known as the *D surface* (for “diamond”).

The application of minimal surfaces to the physical world has so far been descriptive, rather than quantitative. Although explicit analytical equations for the parameters of some TPMSs have recently been derived, problems such as stability and mechanical strength are unresolved. While describing structure using the concept of curvature is mathematically attractive, it has yet to make its full impact on chemistry.

2.7 Enumeration of Crystalline Structures

It is a matter of considerable scientific and practical importance to enumerate all possible networks of atoms in a systematic way. For example, 4-connected networks (that is, networks in which each atom is connected to exactly four neighbors) occur in crystalline elements, hydrates, covalently bonded crystals, silicates, and many synthetic compounds. Of particular interest is the possibility of using systematic enumeration to discover and generate new *nanoporous architectures*.

Nanoporous materials are materials with tiny holes in them that allow some substances to pass through and not others. Many are naturally occurring, such as cell membranes and “molecular sieves” called *zeolites*, but many others have been synthesized. There are now 152 recognized structure types of zeolites, with several new types being added to the list every year. Zeolites find many important applications in science and technology, in areas as diverse as catalysis, chemical separation, water softening, agriculture, refrigeration, and optoelectronics. Unfortunately, the problem of enumeration is fraught with difficulties, and since the number of 4-connected 3D networks is infinite and there is no systematic procedure for their derivation, the results reported so far have been obtained by empirical methods.

Enumeration originated with the work of Wells (1984) on 3D nets and polyhedra. Many possible new structures were found by model building or computer search algorithms. New research in this field is based on recent advances in combinatorial tiling theory, developed by the first generation of pure mathematicians familiar with computing. The tiling approach identified over nine hundred networks with one, two, and three kinds of inequivalent vertices, which we call uninodal, binodal, and trinodal.

However, only a fraction of the mathematically generated networks are chemically feasible (many would be “strained” frameworks requiring unrealistic bond lengths and bond angles), so for the mathematics to be useful an effective filtering process is needed to identify the most plausible frameworks. Methods of computational chemistry were therefore used to minimize the framework energy of the various hypothetical structures, which were treated as though they were made from silicon dioxide. The unit cell parameters, framework energies and densities, volumes available to adsorption, and X-ray diffraction patterns were all

calculated. A total of 887 structures were successfully optimized and ranked according to their framework energies and available volumes to give a subset of chemically feasible hypothetical structures. A number of them have since been synthesized.

The results of these calculations are relevant to the structures of zeolites and other silicates, aluminophosphates (AlPOs), oxides, nitrides, chalcogenides, halides, carbon networks, and even to polyhedral bubbles in foams.

2.8 Global Optimization Algorithms

A wide variety of problems in practically all fields of physical science involve *global optimization*, that is, determining the global minimum (or maximum) of a function of an arbitrary number of independent variables (Wales 2004). These problems also appear in technology, design, economics, telecommunications, logistics, financial planning, travel scheduling, and the design of microprocessor circuitry. In chemistry and biology, global optimization arises in connection with the structure of clusters of atoms, protein conformation, and molecular docking (the fitting and binding of small molecules at the active sites of biomacromolecules such as enzymes and DNA). The quantity to be minimized is nearly always the energy of the system (see below).

Global optimization is like trying to find the deepest point in a very rugged landscape. In most cases of practical interest it is very difficult because of the ubiquity of *local minima*, or holes in the landscape, the number of which tend to increase exponentially with the size of the problem. Conventional minimization techniques are time-consuming and have a tendency to find a nearby hole and stay there: that is, they converge to whichever local minimum they first encounter. The *genetic algorithm* (GA), an approach inspired by Darwin’s theory of evolution, was introduced in the 1960s. This algorithm starts with a set of solutions (represented by “chromosomes”) called a *population*. Solutions from one population are taken and used to form a new population. This is done in such a way that one expects the new population to be better than the old one. Solutions that are chosen for forming new solutions (“offspring”) are selected according to their “fitness”: the more suitable they are the more chances they have to reproduce. This is repeated until some condition is satisfied. (For example, one might stop after a certain number of generations or after a certain improvement of the solution has been achieved.)

Simulated annealing (SA), introduced in 1983, uses an analogy between the annealing process, in which a molten metal cools and freezes into a minimum-energy structure, and the search for a minimum in a more general system. The process can be thought of as an adiabatic approach to the lowest-energy state. The algorithm employs a random search which accepts not only changes that decrease the energy, but also some changes that increase it. The energy is represented by an *objective function* f , and the energy-increasing changes are accepted with a probability $p = \exp(-\delta f/T)$, where δf is the increase in f and T is the system “temperature,” irrespective of the nature of the objective function. SA involves the choice of “annealing schedule,” initial temperature, the number of iterations at each temperature, and the temperature decrease at each step as cooling proceeds.

Taboo (or tabu) search is a general-purpose stochastic global-optimization method originally proposed by Glover in 1989. It is used for very large combinatorial optimization tasks and has been extended to continuous-valued functions of many variables with many local minima. Taboo search uses a modification of “local search,” which starts from some initial solution and attempts to find a better solution. This becomes the new solution and the process restarts from it. The procedure continues step by step until no improvement is found to the current solution. The algorithm avoids entrapment in local minima and gives the optimal final solution. A recent method of global optimization, known as “basin hopping,” has been successfully applied to a variety of atomic and molecular clusters, peptides, polymers, and glass-forming solids. The algorithm is based upon a transformation of the energy landscape that does not affect the relative energies of local minima. Combined with taboo search, basin hopping shows a significant improvement in efficiency over the best published results for atomic clusters.

2.9 Protein Structure

Proteins are linear sequences of amino acids, molecules containing both the amide ($-\text{NH}_2$) and carboxylic ($-\text{COOH}$) functional groups. Understanding the means by which a protein adopts its 3D structure is a key scientific challenge (Wales 2004). This problem is also critical to developing strategies, at the molecular level, to counter “protein folding diseases” such as Alzheimer’s disease and “mad cow” disease. The strategy in tackling protein folding relies upon the fact, observed by Anfinsen, Haber, Sela, and White in 1961, that the structure

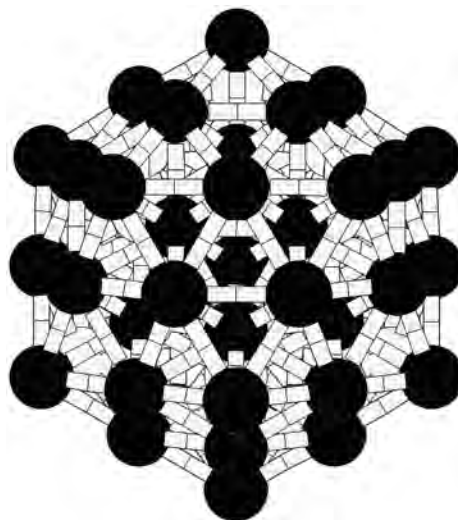


Figure 5 A fifty-five-atom Lennard-Jones cluster.
(Courtesy of Dr. D. J. Wales, Cambridge University.)

of a folded protein corresponds to the conformation which minimizes the free energy of the system. The free energy of a protein depends on the various interactions within the system, and each can be modeled mathematically using the principles of electrostatics and physical chemistry. As a result, the free energy of a protein can be expressed as a function of the positions of the constituent atoms. The 3D arrangement of the protein then corresponds to the set of atomic locations providing the minimum possible value of the free energy, and the problem is reduced to finding the global minimum of the potential-energy surface of the protein. The problem is further complicated because some proteins require other molecules, “chaperones,” to enable them to reach a particular configuration.

2.10 Lennard-Jones Clusters

Lennard-Jones clusters are closely packed arrangements of atoms in which every pair of atoms has an associated potential energy, given by the classical *Lennard-Jones potential-energy function*. The *Lennard-Jones cluster problem* is to determine the atomic cluster configurations with minimum potential energy (figure 5). If n is the number of atoms in the cluster, then one wishes to find points p_1, p_2, \dots, p_n so as to minimize the sum

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^n (r_{ij}^{-12} - 2r_{ij}^{-6}),$$

where r_{ij} stands for the Euclidean distance between p_i and p_j , and the atoms of the cluster are positioned at p_1, p_2, \dots, p_n . The problem is still a challenge, both to optimization methods and to computer technology. A systematic survey by Northby in 1987, which yielded most of the lowest Lennard-Jones potential values in the range $13 \leq n \leq 147$, was a significant landmark, and these results have since been improved by about 10%. The results for $n = 148, 149, 150, 192, 200, 201, 300$, and 309 have now been reported using stochastic global-optimization algorithms.

2.11 Random Structures

Stereology, originally the deduction of 3D structure from microscope examination of sections, has required the development of a substantial branch of statistical mathematics, in which R. E. Miles and R. Coleman have played leading roles. Stereology concerns the estimation of geometrical quantities. Geometrical shapes are used to probe objects to learn about their quantities, such as volume or length. Random sampling is a basic step in all stereological estimation. The degree of randomness required for any estimate varies.

Even apparently simple questions involving randomness with spatial constraints may prove difficult. For example, Gotoh and Finney gave an estimate of 0.6357 as the density expected for a dense random packing of hard spheres of equal size, and their answer to this apparently simple question has not since been improved upon, as far as we know. The problem needs to be defined very carefully, since it is far from obvious what one means by a “random packing” of spheres. This is even more true when one investigates other, related problems concerning the interaction of molecules using computer simulation. This area, called *molecular dynamics*, was begun by A. Rahman, and it developed steadily from the 1960s as computers themselves developed. An example of a problem in molecular dynamics is the modeling of liquid water. This is still difficult, but the immense computing power that is now available has enabled enormous progress to be made.

3 Process

In 1951 Belousov discovered the *Belousov-Zhabotinski reaction*, in which time-dependent spatial patterns appear in an apparently isotropic medium. The mechanism of this reaction was elucidated in 1972, and this opened up an entire new research area: *nonlinear*

chemical dynamics. Oscillatory phenomena have also been observed in *membrane transport*. Winfree and Prigogine have shown how patterns in space and time can appear, and some of these patterns have been fitted to practical examples.

The development of *cellular automata* began with Stanisław Ulam, Lindenmeyer systems, and Conway’s “game of life” and continues to this day. With his huge book, Wolfram (2002) has demonstrated the complexity that can arise from apparently simple rules, and recently Reiter has used cellular automata to simulate the growth of snowflakes, beginning to answer questions that Kepler posed in 1611. There is a group of mathematicians in Bielefeld, led by Andreas Dress, who deal with structure-forming processes; they have made particular progress in modeling actual chemistry and thus revealing possible mechanisms.

4 Search

4.1 Chemical Informatics

A fundamental development in chemistry has been the application of computing to searching multidimensional databases of chemical compounds and their structures. These databases are now enormous compared with their (already large) predecessors, the classical Gmelin and Beilstein databases. The search process has required fundamental mathematical analyses, as exemplified in the pioneering work of Kennard and Bernal in developing the Cambridge Structural Database (www.ccdc.cam.ac.uk/products/csd/).

What is the best way to encode the structure of a 3D molecule or a crystal arrangement as a linear sequence of symbols? One would like to be able to restore the structure efficiently from its encoding, and also to search efficiently through a big list of encoded structures. The problems that this raises are of long standing, and need insights both from mathematics and chemistry.

4.2 Inverse Problems

Many of the mathematical challenges of chemistry are inverse problems. Often they involve solving a set of linear equations. If there are as many equations as unknowns and the equations are independent, then this can be done by inverting a square matrix. However, if the system is singular or redundant, or if there are fewer equations or more equations than unknowns, then the corresponding matrix is singular or rectangular and there is no ordinary inverse. Nevertheless, it

is possible to define a *generalized inverse*, which gives a good model for linear problems. (It is the so-called *Moore–Penrose inverse* or *pseudo-inverse* involved in singular value decomposition.) This always exists and it uses all available information; it is related to the problem of reconstructing a 3D structure from a 2D projection. The operation has been fully described and is now available in Mathematica.

The generalized inverse also enables one to handle redundant axes in quasicrystals, but usually the interesting problems are nonlinear. Other inverse problems include the following.

- (i) Finding the arrangement of atoms that gives rise to the observed scattering patterns of X-rays or electrons from a crystal.
- (ii) Reconstructing a 3D image from 2D projections in microscopy or X-ray tomography.
- (iii) Reconstructing the geometry of a molecule given probable interatomic distances (and perhaps bond angles and torsion angles).
- (iv) Finding the way in which a protein molecule folds to give an active site, given the sequence of constituent amino acids.
- (v) Finding the pathway to producing a molecule synthetically, given that it occurs in nature.
- (vi) Finding the sequence of rules that generate a membrane or a plant or another biological object, given that it takes a certain shape.

Some questions of this type do not have unique answers. For example, the classic question as to whether the shape of a drumhead can be determined from its vibration spectrum (can you hear the shape of a drum?) has been answered in the negative: two vibrating membranes with different shapes may have the same spectrum. It was thought that this ambiguity might also be the case for crystal structures. Linus Pauling suggested that there might be two different crystal structures that were *homometric* (that is, giving the same diffraction pattern), but no definite example has been found.

5 Conclusion

As the examples in this article show, mathematics and chemistry have a symbiotic relationship, with developments in one often stimulating advances in the other. Many interesting problems, including several that we have mentioned here, are still waiting to be solved.

Further Reading

- Cotton, F. A. 1990. *Chemical Applications of Group Theory*. New York: Wiley Interscience.
- Hollas, J. M. 2003. *Modern Spectroscopy*. New York: John Wiley.
- Hyde, S., S. Andersson, K. Larsson, Z. Blum, T. Landh, S. Lidin, and B. W. Ninham. 1997. *The Language of Shape. The Role of Curvature in Condensed Matter: Physics, Chemistry and Biology*. Amsterdam: Elsevier.
- Parr, R. G., and W. Yang. 1989. *Density-Functional Theory of Atoms and Molecules*. Oxford: Oxford University Press.
- Thomas, J. M. 2003. Poetic suggestion in chemical science. *Nova Acta Leopoldina* NF 88:109–39.
- Wales, D. J. 2004. *Energy Landscapes*. Cambridge: Cambridge University Press.
- Wells, A. F. 1984. *Structural Inorganic Chemistry*. Oxford: Oxford University Press.
- Wolfram, S. 2002. *A New Kind of Science*. Champaign, IL: Wolfram Media.

VII.2 Mathematical Biology

Michael C. Reed

1 Introduction

Mathematical biology is an extremely large and diverse field. The objects of study range from molecules to global ecosystems and the mathematical methods come from many of the subdisciplines of the mathematical sciences: ordinary and partial differential equations, probability theory, numerical analysis, control theory, graph theory, combinatorics, geometry, computer science, and statistics. The most that one short article can do is to illustrate by selected examples this diversity and the range of new mathematical questions that arise naturally in the biological sciences.

2 How Do Cells Work?

From the simplest point of view, cells are large biochemical factories that take inputs and manufacture lots of intermediate products and outputs. For example, when a cell divides, its DNA must be copied and that requires the biochemical synthesis of large numbers of adenine, cytosine, guanine, and thymine molecules. Biochemical reactions are usually catalyzed by enzymes, proteins that facilitate a reaction but are not used up by it. Consider, for example, a reaction in which chemical A is converted to chemical B with the help of an enzyme E. If $a(t)$ and $b(t)$ are the respective concentrations of A and B at time t , then one typically writes down a differential equation for $b(t)$, which

takes the form

$$b'(t) = f(a, b, E) + \dots - \dots$$

Here, f is the rate of production, which typically depends on a , b , and E . Of course B may be produced by other reactions (which would lead to additional positive terms $+\dots$) and may be used as a substrate itself in still other reactions (which would lead to additional negative terms $-\dots$). So, given a particular cell function or biochemical pathway, we can just write down the appropriate set of nonlinear coupled ordinary differential equations for the chemical concentrations and solve it by hand or by machine computation. However, this straightforward approach is often unsuccessful. First of all, there are a lot of parameters (and variables) in these equations and measuring them in the context of real living cells is difficult. Second, different cells behave differently and may have different functions, so we would expect the parameters to be different. Third, cells are alive and change what they are doing, so the parameters may themselves be functions of time. But the greatest difficulty is that the particular pathway under study is not really isolated. Rather, it is embedded in a much larger system. How do we know that our model system will continue to behave in the same way when embedded in this larger context? We need new theorems in dynamical systems that answer questions such as this, not for general “complex systems” but for the particular kinds of complex systems that arise in important biological problems.

Cells continue to accomplish many basic tasks even though their environments (i.e., their inputs) are constantly changing. A brief example of this phenomenon, which is known as *homeostasis*, will illustrate the problem of “context.” Let us suppose that the chemical reaction above is one step in the pathway for making the thymines necessary for cell division. If the cell is a cancer cell, we would like to turn off this pathway, and a reasonable way to try to do this would be to put into the cell a compound X that binds to E , thereby reducing the amount of free enzyme available to make the reaction run. Two homeostatic mechanisms immediately come into play. First, a typical reaction is inhibited by its product: that is, f decreases as b increases. This makes biological sense because it ensures that B is not overproduced. So, when the amount of free E is reduced and the rate f declines, the resulting decrease in b drives the rate up again. Second, if the rate f is lower than usual, the concentration a typically rises since A is not being used up as quickly, which also drives the

rate f up again since f increases as a increases. Given the network in which A and B are embedded, one can imagine calculating how much f will drop if we put a certain amount of X into the cell. In fact, f may drop even less than we calculate because of another homeostatic mechanism that is not even in our network. The enzyme E is a protein produced by the cell via instructions from a gene. It turns out that sometimes the concentration of free E inhibits the messenger RNA that codes for the production of E itself. Then, if we introduce X and reduce free E , the inhibition is removed and the cell automatically increases its rate of production of E , thus raising the amount of free E and with it raising the reaction rate f .

This illustrates a fundamental difficulty in studying cell biochemistry, indeed a difficulty in studying many biological systems. These systems are very large and very complex. To gain understanding, it is natural to concentrate on particular relatively simple subsystems. But one always has to be aware that the subsystems exist in a larger context that may contain variables (excluded by the simplification) that are crucial for understanding the behavior and biological function of the subsystem itself.

Although cells exhibit remarkable homeostasis, they also undergo spectacular changes. For example, cell division requires unzipping of the DNA, synthesis of two new complementary strands, the movement apart of the two new DNAs, and the pinching off of the mother cell to produce two daughters. How does a cell do all this? In the case of yeast cells, which are comparatively simple, the actions of the biochemical pathways are quite well understood, partly because of the mathematical work of John Tyson. But as our brief discussion makes clear, biochemistry is not all there is to cell division; an important additional feature is motion. Materials are being transported all the time throughout cells from one specific place to another (so their motion is not just diffusion), and indeed, cells themselves move. How does this happen? The answer is that materials are transported by special molecules called molecular motors that turn the energy of chemical bonds into mechanical force. Since bonds are formed and broken stochastically (that is, some randomness is involved), the study of molecular motors leads naturally to new questions in STOCHASTIC ORDINARY AND PARTIAL DIFFERENTIAL EQUATIONS [IV.24]. A good introduction to the mathematics of cell biology is Fall et al. (2002).

3 Genomics

To understand the mathematics that was involved in sequencing the human genome it is useful to start with the following simple question. Suppose that we cut up a line segment into smaller segments and are presented with the pieces. If we are told the order in which the pieces came in the original segment, then we can put them back together and reconstruct the segment. In general, since there are many possible orders, we cannot reconstruct the segment without extra information of this kind. Now suppose that we have cut up the segment in *two different ways*. Think of the line segment as an interval I of real numbers, and let the pieces be A_1, A_2, \dots, A_r when you cut it up the first way, and B_1, B_2, \dots, B_s when you cut it up the other way. That is, the sets A_i form a partition of the interval I into subintervals, and the sets B_j form another partition. For simplicity, assume that no A_i shares an endpoint with any B_j , except for the two endpoints of I itself.

Suppose that we know nothing about the order in which the pieces A_i and B_j come in I . In fact, suppose that all we know about them is which A_i overlap with which B_j : that is, which of the intersections $A_i \cap B_j$ are nonempty. Can we use this information to work out the original order of the pieces A_i and thereby reconstruct the interval I (or its reflection)? The answer will sometimes be yes and sometimes no. If it is yes, then we would like to find an efficient algorithm for doing the reconstruction, and if it is no, then we would like to know how many different reconstructions are consistent with the given information. This so-called *restriction mapping problem* is really a problem in GRAPH THEORY [III.34]: the vertices of the graph correspond to the sets A_i or B_j , and there is an edge between A_i and B_j if $A_i \cap B_j \neq \emptyset$.

A second problem is whether we can find the original order of the A_i (or the B_j) if what we are told is the length of each set A_i and each set B_j , and the set of all the lengths of the intersections $A_i \cap B_j$. The catch is that we are not told which length corresponds to which intersection. This is called the *double digest problem*. Again one would like to be able to tell when there is only one solution, or to place an upper bound on the number of possible reconstructions if there is more than one.

Human DNA is, for our purposes here, a word of length approximately 3×10^9 over a four-letter alphabet A, G, C, T. That is, it is a sequence of length 3×10^9 in which each entry is A, G, C, or T. In the cell, the word is bound letter by letter to the “complementary”

word, which is determined by the rule that A can only be bound to T, and C can only be bound to G. (For example, if the word is ATTGATCCTG, then the complementary word is TAACTAGGAC.) In this brief discussion we will ignore the complementary word.

Since DNA is so long (it would be approximately two meters if one stretched it out into a straight line) it is very hard to handle experimentally, but the sequence of letters in short segments of approximately five hundred letters can be determined by a process called gel chromatography. There are enzymes that cut DNA wherever specific very short sequences occur. So if we digest a DNA molecule with one of these enzymes and digest another copy with a different enzyme, we can hope to determine which fragments from the first digestion overlap fragments from the second digestion and then use techniques from the restriction mapping problem to reconstruct the original DNA molecule. The interval I corresponds to the whole DNA word, and the sets A_i to the fragments. This involves sequencing and comparing the fragments, which has its own difficulties. However, *lengths* of fragments are not so hard to determine, so another possibility is to digest with the first enzyme and measure lengths, digest with the second and measure lengths, and finally digest with both and measure lengths. If one does this, then the problem one obtains is essentially the double digest problem.

To completely reconstruct the DNA word one takes many copies of the word, digests with enzymes, and selects at random enough fragments that together they have a high probability of covering the word. Each of the fragments is cloned, in order to get enough mass, and then sequenced by gel chromatography. Both processes can introduce errors, so one is left with a very large number of sequenced fragments with known error rates for the letters. These need to be compared to see if they overlap: that is, to see if the sequence near the end of one fragment is the same as (or very similar to) the sequence at the beginning of another. This alignment problem is itself difficult because of the large number of possibilities involved. So, in the end we have a very large restriction mapping problem except that we can only say that given fragments overlap with probabilities that are themselves hard to estimate. A further difficulty is that DNA tends to have large blocks that repeat in different parts of the word. As a result of these complications, the problem is much harder than the restriction mapping problem described earlier. It is clear that graph theory, combinatorics, probability

Terri: ‘ \emptyset ’
chosen for
this symbol
and used
globally
throughout
Companion.
OK?

theory, statistics, and the design of algorithms all play central roles in sequencing a genome.

Sequence alignment is important in other problems as well. In phylogenetics (see below) one would like a way of saying how similar two genes or genomes are. When studying proteins, one can sometimes predict protein three-dimensional structure by searching databases for known proteins with the most similar amino acid sequence. To illustrate how complex these problems are, consider a sequence $\{a_i\}_{i=1}^{1000}$ of one thousand letters from our four-letter alphabet. We wish to say how similar it is to another sequence $\{b_i\}_{i=1}^{1000}$. Naively, one could just compare a_i with b_i and define a METRIC [III.58] like $d(\{a_i\}, \{b_i\}) = \sum \delta(a_i, b_i)$. However, DNA sequences have evolved typically by insertions and deletions as well as by substitutions. Thus if the sequence ACACAC... lost its first C to become AACAC..., the two sequences would be very far apart in this metric even though they are very similar and related in a simple way. The way around this difficulty is to allow sequences to include a fifth symbol, -, which stands for the place of a deletion or a place opposite an insertion. Thus, given two sequences (of perhaps different lengths), we wish to find how they can be augmented with dashes to give the minimum possible distance between them. A little thought will convince the reader that it is not feasible to use a brute-force search for a problem like this, even for the fastest computers—there are so many potential augmentations that the search would take far too long. Serious and thoughtful algorithm development is required. Two excellent introductions to the material discussed in this section are Waterman (1995) and Pevzner (2000).

4 Correlation and Causality

The central dogma of molecular biology is DNA \rightarrow RNA \rightarrow proteins. That is, information is stored in DNA, it is transferred out of the nucleus by RNA, and the RNA is then used in the cell to make proteins that carry out the work of the cell through the metabolic processes discussed in section 2. Thus DNA directs the life of the cell. Like most things in biology, the true situation is much more complicated. Genes, which are segments of DNA that code for the manufacture of particular proteins, are sometimes turned on and sometimes turned off. Usually, they are partially turned on; that is, the protein they code for is manufactured at some intermediate rate. This rate is controlled by the binding (or lack of binding) of small molecules or specific proteins

to the gene, or to the RNA that the gene codes for. Thus genes can produce proteins that inhibit (or excite) other genes; this called a gene network.

In a way, this was obvious all along. If cells can respond to their environments by changing what they do, they must be able to sense the environment and signal the DNA to change the protein content of the cell. Thus, while sequencing DNA and understanding specific biochemical reactions are important first steps in understanding cells, the hard and interesting work to come is to understand *networks* of genes and biochemical reactions. It is these networks, in which proteins control genes and genes control proteins, that carry out and control specific cellular functions. The mathematics will be ordinary differential equations for chemical concentrations and variables that indicate to what extent a gene is turned on. Since transport into and out of the nucleus occurs, partial differential equations will be involved. And, finally, since some of the molecular species occur in very small numbers, concentration (molecules per unit volume) may not be a useful approximation for computations about chemical binding and dissociation: they are probabilistic events.

Two kinds of statistical data can give hints about the components of these gene networks. First, there are large numbers of population studies that correlate specific genotypes to specific phenotypes (such as height, enzyme concentration, cancer incidence). Second, tools known as *microarrays* allow us to measure the relative amounts of a large number of different messenger RNAs in a group of cells. The amount of RNA tells us how much a particular gene is turned on. Thus, microarrays allow us to find correlations that may indicate that certain genes are turned on at the same time or perhaps in a sequence. Of course, correlation is not causality and a consistent sequential relationship is not necessarily causal either (sure, football causes winter, a sociologist once said). Real biological progress requires understanding the gene networks discussed above; they are the mechanisms by which the genotypes play out in the life of the cell.

A nice discussion of the relationship between population correlations and mechanisms occurs in Nijhout (2002), from which we take the following simple example. Most phenotypic traits depend on many genes; suppose that we consider a trait that depends on only two genes. Figure 1 depicts a surface that shows how the trait in an individual depends on how much each of the genes is turned on. All three variables are scaled from 0 to 1. Suppose that we study a population whose

Terri: proofreader thought the correlation between the text here and the figure was suspect, but Tim is certain that it's all OK and that the figure shows exactly what it is meant to show. OK?

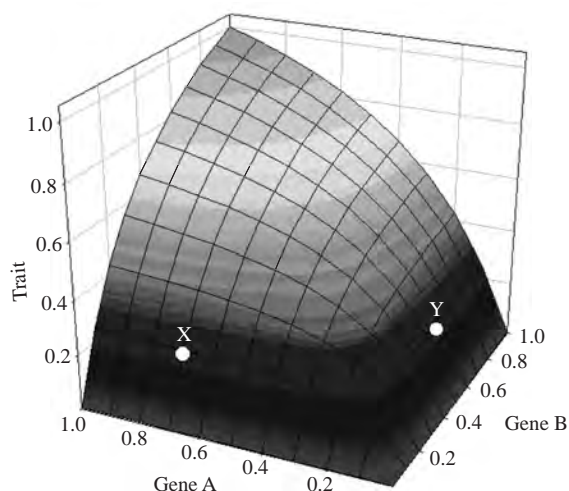


Figure 1 A phenotypic surface.

members have a genetic makeup that puts the individuals near the point X on the graph. If we do a statistical analysis of the population, we will find that gene B is highly statistically correlated to the trait, but gene A is not. On the other hand, if the individuals in the population all live near the point Y on the surface, we will discover in our population study that gene A is highly statistically correlated to the trait, but gene B is not. More detailed examples with specific biochemical mechanisms are discussed in Nijhout's paper. Similar examples can be given for microarray data. This does not mean that population studies or microarray data are unimportant. Indeed, in studying hugely complex biological systems, statistical information can suggest where to look for the mechanisms that will ultimately give biological understanding.

5 The Geometry and Topology of Macromolecules

To illustrate the natural geometric and topological questions that arise when one studies macromolecules, we will briefly discuss molecular dynamics, protein-protein interactions, and the coiling of DNA. Genes code for the manufacture of proteins, which are large molecules made up of sequences of amino acids. There are twenty amino acids, each coded by a triplet of base pairs, and a typical protein might have five hundred amino acids. Interactions among the amino acids cause the protein to fold up into a complicated three-dimensional shape. This three-dimensional structure is

crucial for the function of the protein since the exposed groups and the nooks and crannies in the shape govern the possible chemical interactions with small molecules and other proteins. Three-dimensional structures of proteins can be approximately determined by X-ray crystallography and nontrivial inverse scattering calculations. The forward problem—namely, given the sequence of amino acids, predict the three-dimensional structure of the protein—is important not only for understanding existing proteins, but also for the pharmacological design of new proteins to accomplish specific tasks. Thus, in the past twenty years a large field called *molecular dynamics* has arisen, in which one uses classical mechanical methods.

Suppose we have a protein that consists of N atoms. Let \mathbf{x}_i denote the position (specified by three real coordinates) of the i th atom, and let \mathbf{x} denote the vector formed from all these coordinates (which belongs to \mathbb{R}^{3N}). For each pair of atoms, one attempts to write down a good approximation to the potential energy, $E_{i,j}(\mathbf{x}_i, \mathbf{x}_j)$, due to their pairwise interaction. This could be the electrostatic interaction, for example, or the van der Waals interaction, which is a classical mechanical formulation of quantum effects. The total potential energy is $E(\mathbf{x}) \equiv \sum E_{i,j}(\mathbf{x}_i, \mathbf{x}_j)$ and Newton's equations of motion take the form

$$\dot{\mathbf{v}} = -\nabla E(\mathbf{x}), \quad \dot{\mathbf{x}} = \mathbf{v},$$

where \mathbf{v} is the vector of velocities. Starting with some initial conditions one can try to solve these equations to follow the dynamics of the molecule. Note that this is a very high-dimensional problem. A typical amino acid has twenty atoms, so that is sixty coordinates right there, and if we are looking at a protein made up of five hundred amino acids, then \mathbf{x} will be a vector with thirty thousand coordinates. Alternatively, one could assume that the protein will fold to the configuration that has the minimum potential energy. Finding this configuration would mean finding the roots of $\nabla E(\mathbf{x})$, by NEWTON'S METHOD [II.4 §2.3] say, and then checking to see which root gives the lowest energy. Again this is an enormous computational task.

It is not surprising that molecular dynamics calculations have been only moderately successful and have predicted the shapes of only relatively small molecules and proteins. The numerical problems are substantial and the choice of energy terms is somewhat speculative. Even more importantly, context matters, as it does in many biological problems. The way proteins fold depends on properties of the solution in which

they sit. Many proteins have several preferred configurations and switch from one to the other depending on interactions with small molecules or other proteins. Finally, it has recently been discovered that proteins do not fold up by themselves from their linear configuration to their three-dimensional shape, but are helped and guided by other proteins called chaperones. It is natural to ask whether there are quantifiable geometrical units larger than points (atoms) that could reasonably form the basis for a good approximation to the dynamics of large molecules.

A start has been made in this direction by groups studying the interactions of proteins with small molecules and other proteins. These interactions are fundamental to cell biochemistry, cell-transport processes, and cell signaling, and so progress is vital to understanding how cells work. Suppose one has two large proteins that are bound to each other. The first thing one would like to do is describe the geometry of the binding region. One could do this as follows. Consider an atom in either protein that is at point x . Given another atom at point y , there is a plane that divides \mathbb{R}^3 into two open half-spaces: the points closer to x and the points closer to y . Now let R_x denote the intersection of all such open half-spaces as y ranges over the positions of all other atoms: that is, R_x consists of those points that are closer to x than to any other atom. The union of the boundaries, $\bigcup_x \partial(R_x)$, called a *Voronoi surface*, consists of triangles and pieces of planes and has the property that each point on the surface is equidistant from at least two atom positions. To model the binding region between the two proteins, we discard all pieces of the Voronoi surface that are equidistant from two atoms that belong to the same protein and keep just the ones that are equidistant from two atoms that are in different proteins. This surface goes off to infinity, so we clip off the parts that are not “close” to either protein. The result is a surface with a boundary made up of polyhedral faces that is a reasonable approximation of the interaction interface between the two proteins. (This is not quite an accurate description: in the actual construction, “distance” is weighted in a way that depends on the atoms involved.) Now choose colors representing the twenty amino acids and color each side of each polyhedral piece with the color of the amino acid that the closest atom is in. This divides each side of the surface into large colored patches corresponding to nearness of a particular amino acid on that side. The coloring of the two sides of the boundary surface will be different, of course, and the placement of

the patches gives information about which amino acids in one protein are interacting with which amino acids in the other. In particular, one amino acid in one protein may interact with several in the other. This gives a way of using geometry to classify the nature of the particular protein-protein interaction.

Finally, let us touch on questions involving the packaging of DNA. The basic problem is easy to see. As mentioned earlier, the human DNA double helix when stretched out linearly is about two meters long. A typical cell has a diameter of about one-hundredth of a millimeter and its nucleus has a diameter of about one-third that size. All of that DNA has to be packed into the nucleus. How is this done?

At least the first stages are well understood. The DNA double helix is wound around proteins called *histones*, which consist of about two hundred base pairs each, yielding chromatin, which is a sequence of such DNA-wrapped histones connected by short segments of DNA. Then the chromatin is itself wrapped up and compacted; the geometrical details are not completely understood. It is important to understand the packing and the mechanisms that create it, because the life of the cell requires unpacking! When the cell divides, the entire DNA helix must be unzipped to form two separate strands, which are the templates on which the two new copies of DNA will be built. Clearly this cannot be done all at once but must involve local unwinding of the DNA off the histones, local unzipping, synthesis, and then local repacking.

It is equally challenging to understand the sequence of events that occurs when a protein is synthesized from a gene. Transcription factors diffuse into the nucleus and bind to specific short segments of DNA (of about ten base pairs) in the regulatory region of the gene. Of course, they will randomly bind wherever they see the same segment. Typically, one needs the binding of several different transcription factors in the regulatory region along with RNA polymerase to start transcription of a gene. That process involves the unwinding of the gene-coding region from the histones so that it can be transcribed, the transport of the resulting RNA out of the nucleus, and the recompactification of the DNA. To understand these processes fully, one will have to solve problems in partial differential equations, geometry, combinatorics, probability theory, and topology. DeWitt Sumners is the mathematician who brought the topological problems in the study of DNA (links, twists, knots, supercoiling) to the attention of the mathematics community. A good reference for molecular

dynamics and the general mathematical issues posed by biological macromolecules is Schlick (2002).

6 Physiology

When one first studies human physiological systems, they seem almost miraculous. They accomplish enormous numbers of tasks simultaneously. They are robust but capable of quick changes if the situation warrants. They are made up of large numbers of cells that actively cooperate so that the tasks of the whole can be done. It is the nature of many of these systems that they are complex, controlled by feedback, and integrated with each other. It is the job of mathematical physiology to understand how they work. We will illustrate some of these points by discussing problems in biological fluid dynamics.

The heart pumps blood throughout a circulatory system that consists of vessels of diameter as large as 2.5 cm (the aorta) and as small as 6×10^{-4} cm (the capillaries). Not only are the vessels flexible, but many are surrounded by muscle and can contract to exert local force on the blood. The main force-generating mechanism (the heart!) is approximately periodic, but the period can change. The blood itself is a very complicated fluid. About 40% of its volume is made up of cells: red blood cells carry most of the oxygen and CO_2 ; white blood cells are immune system cells that hunt bacteria; and platelets are part of the blood clotting process. Some of these cells have diameters that are larger than the smallest capillaries, which raises the nice question of how they get through. You notice that we are very far away from most of the simplifying assumptions of classical fluid dynamics.

Here is an example of a circulatory-system question. In a significant number of people, the mitral valve, which is the inflow valve to the left side of the heart, becomes defective. It is common to replace the valve by an artificial one and this leads to an important question: how should one design the artificial valve so that the resulting flow in the left heart chamber has as few stagnant points as possible, since clots tend to form at these points? Charles Peskin did the pioneering work on this problem. Here is another question. The white blood cells are not carried in the middle of the fluid but tend to roll along the walls. Why do they do that? It is a good thing that they do, because their job is to sniff out inflammation outside the blood vessel and, when they find it, to stop and burrow through the blood vessel wall to get to the inflamed site. Another circulatory fluid dynamics question is discussed in section 10.

The circulatory system is connected to many other systems. The heart has its own pacemaker cells, but its frequency of contraction is regulated by the autonomic nervous system. Through the *baroreceptor reflex*, the sympathetic nervous system tightens blood vessels to avoid a dramatic drop in blood pressure when we stand. Overall average blood pressure is maintained by a complicated regulatory feedback mechanism involving the kidneys. It is worthwhile remembering that all these things are being accomplished by living tissues whose parts are always decaying and being replaced. For example, the gap junctions that transmit current at very low resistance between heart muscle cells have a half-life of approximately one day.

As a final example, we consider the lung, which has a fractal branching structure that terminates after twenty-three levels in about 600 million air sacs called *alveoli*, in which oxygen and CO_2 are exchanged with the circulating blood. The Reynolds number of the air flow varies by about three orders of magnitude between the large vessels near the throat and the tiny vessels near the alveoli. Premature infants often have respiratory difficulty because they lack surfactants that reduce surface tension on the inner surfaces of the alveoli. The high surface tension makes the alveoli collapse, which makes breathing difficult. One would like the infants to breathe in air that includes tiny aerosol drops of surfactant. How small should the drops be so that as much surfactant as possible makes it to the alveoli?

The mathematics of physiology consists mostly of ordinary and partial differential equations. However, there is a new feature: many of these equations have time delays. For example, the rate of respiration is controlled by a brain center that senses the CO_2 content of blood. It takes almost fifteen seconds for blood to go from the lungs to the left heart and from there to the brain center. This time delay is even longer in patients with weak hearts and often these patients display Cheyne–Stokes breathing: very rapid breathing alternates with periods of little or no breathing. Such oscillations in control systems are well-known as the time delay gets longer. Since partial differential equations are often involved, new mathematical results are needed that go well beyond the standard theory of ordinary differential equations with delay, which was initiated by Bellman in the 1950s. An excellent reference for the applications of mathematics to physiology is Keener and Sneyd (1998).

Terri: Tim thinks this is clear. (And for what it's worth, I do too!)

7 What's Wrong with Neurobiology?

The short answer is that there is not enough theory. This may seem an odd thing to say, since neurobiology is the home of the Hodgkin-Huxley equations, which are often cited as a triumph of mathematics in biology. In a series of papers in the early 1950s, Hodgkin and Huxley described several experiments, and gave a theoretical basis for explaining them. Building on the work of physicists and chemists (for example, Walter Nernst, Max Planck, and Kenneth Cole), they discovered the relationship between certain ionic conductances and the trans-membrane electrical potential, $v(x, t)$, in the axons of neurons, and they formulated a mathematical model:

$$\begin{aligned}\frac{\partial v}{\partial t} &= \alpha \frac{\partial^2 v}{\partial x^2} + g(v, y_1, y_2, y_3), \\ \frac{\partial y_i}{\partial t} &= f_i(v, y_i), \quad i = 1, 2, 3.\end{aligned}$$

Here the y_i are related to the membrane conductances of various ions. The equations have solutions that are pulses that keep their shape and travel at constant velocity in a way that corresponds to the observed behavior of action potentials in real neurons. The ideas, both explicit and implicit, in these discoveries form the basis of much single-neuron physiology. Of course, mathematicians should not be too proud about this since Hodgkin and Huxley were biologists. The Hodgkin-Huxley equations were part of the stimulus for interesting work by mathematicians on traveling waves and pattern formation in reaction-diffusion equations.

However, not everything can be explained at the level of just one neuron. Watch your hand as it reaches out gracefully to pick up an object. Think about the so-called ocular-vestibular reflex in which motions of the head are automatically compensated for by motions of the eyes so that your gaze can remain fixed. Consider the fact that you are looking at stereotypical black marks on a page and they mean something inside your head. These are *system properties*, and the systems are large indeed. There are approximately 10^{11} neurons in the central nervous system and on average each makes about one thousand connections to other neurons. These systems will not be understood just by examining their parts (the neurons) and, for obvious reasons, experimentation is limited. Thus, experimental neurobiology, like experimental physics, needs input from deep and imaginative theorists.

The lack of a large theory community interacting robustly with experimentalists is to some extent a historical accident. Grossberg asked how groups of (quite simple) model neurons, if they were connected in the right ways, could accomplish various tasks such as pattern recognition and decision making, or could exhibit certain “psychological” properties (Grossberg 1982). He also asked how these networks could be trained. At about the same time it was shown that networks of neuron-like elements connected in the right way could automatically compute good solutions of large, difficult problems like the TRAVELING-SALESMAN PROBLEM [VII.5 §2]. These and other factors, including the great interest in software engineering and artificial intelligence, led to the emergence of a large community of researchers studying “neural networks.” The members of this community were mostly computer scientists and physicists, so it was natural for them to concentrate on the design of devices, rather than biology. This was noticed, of course, by experimental neurobiologists, who lost interest in collaborating with these theorists.

This brief history is of course an oversimplification. There are mathematicians (and physicists and computer scientists) who are essentially theoreticians for neuroscience. Some of them work on hypothetical networks, typically either very small networks or networks with strong homogeneity properties, to discover what are the emergent behaviors of the systems. Others work on modeling real physiological neural networks, often collaboratively with biologists. Usually, the models consist of ordinary differential equations for the firing rates of the individual neurons or mean-field models that involve integral equations. These mathematicians have made real contributions to neurobiology.

But much more is needed, and to see why, it is useful to think about just how difficult these problems really are. First, there is no one-to-one correspondence between the cells of the central nervous system in different members of the same species (except in special cases like *C. elegans*). Second, neurons in the same animal differ widely in their anatomy and physiology. Third, the details of a particular network may well depend on the life history of the animal. Fourth, most neurons are somewhat unreliable devices in that they give different outputs under repeated trials with the same input. Finally, one of the prime characteristics of neural systems is that they are plastic, adaptable, and ever changing. After all, if you remember anything of what is written here, then your head is different from when you began. Between the level of the single neuron

Terri: it's Tim's opinion (and mine) that this abbreviation is clear enough and that including the full first word wouldn't help readers and would, perhaps, look strange to a biologist. OK as it is?

and the psychological level, there are probably twenty levels of networks, each network feeding into and being controlled by networks at other levels. The mathematical objects that will enable us to classify, analyze, and understand how this all works have probably not yet been discovered.

8 Population Biology and Ecology

Let us begin with a simple example. Imagine a large orchard of equally spaced trees and suppose that one tree has a disease. The disease can be transmitted only to nearest neighbors, and is transmitted with probability p . What is $E(p)$, the expected percentage of trees that will be infected? Intuitively, if p is small, $E(p)$ should be small, and if p is large, $E(p)$ should be close to 100%. In fact, one can prove that $E(p)$ changes very rapidly from being small to being large as p passes through a small transition region around a particular critical probability p_c . One would expect p to decrease as the distance, d , between trees increases; farmers should choose d in such a way that p is less than the critical probability, in order to make $E(p)$ small. We see here a typical issue in ecological problems: how does behavior on the large scale (tree epidemic or not) depend on behavior at the small scale (the distance between trees). And, of course, the example illustrates that understanding the biological situation requires mathematics. For other examples of sharp global changes in probabilistic models, see PROBABILISTIC MODELS OF CRITICAL PHENOMENA [IV.25].

Suppose that we now widen our gaze to consider forests—let us say the forests on the East coast of the United States. We would like to understand how they have come to be as they are. Most of them were not planted in neat rows, so that is already a complication. But there are two other really new features. First, there is not one species but many, and each species of tree has different properties: shape, seed dispersal, need for light, and so forth. The species are different, but their properties affect each other because they are living in the same space. Second, the species, and the interactions between the species, are affected by the physics of the environment. There are physical parameters that vary on long timescales, like average temperature, and there are other parameters that vary on very short timescales, like wind speed (for seed dispersal). Certain properties of forests may depend on the fluctuations in these parameters as much as on the values themselves. Finally, one might have to take into account

the reaction of the ecosystem to catastrophic events such as hurricanes or prolonged drought.

The difficulties are similar to those we have seen for other problems in mathematical biology. One would like to understand the emergent behavior on the large scale. To do this one creates mathematical models that relate the behavior on the small scale to the large scale. However, on the small scale one is overwhelmed by the biological details. Which of these details should be in the model? Of course, there is no simple answer to this because, in fact, this is the heart of what we want to know. Which of the bewildering variety of local properties or variables give rise to the large-scale behavior and by what mechanisms? Furthermore, it is not obvious what kinds of model are best. Should we model each individual and its interactions, or should we use population densities? Should we use deterministic models or stochastic models? These are also hard questions, and the answers depend on the system being studied and the questions being asked. A nice discussion of these different modeling choices can be found in Durrett and Levin (1994).

Let us focus again on a simple model: the so-called *SIRS model* for the spread of a disease in a population. A crucial parameter is the *infectious contact number*, σ , which represents the average number of new infections that an infected individual creates in the susceptible population. For a serious disease one would like to bring the value of σ down to below 1 (so that an epidemic will be unlikely) by vaccination, which takes individuals from the susceptible category and puts them in the removed category. Since vaccination is expensive and it is difficult to vaccinate high percentages of the population, it is an important public-health problem to know how much vaccination is needed to bring σ to below 1. A little reflection shows us how difficult this problem really is. First of all, the population is not well mixed, so one may not be able to ignore spatial separation, as is done in the SIRS model. Even more important, σ depends on the social behavior of individuals and the subclasses of the population to which they belong (as anyone with small children in school will attest). Thus, we see a genuinely new issue here: if an ecological problem involves animals, then the social behavior of the animals may affect the biology.

In fact, the issues are even deeper. Individuals in groups, or species, or subpopulations, vary and it is just this variation on which natural selection acts. So, to understand how an ecosystem got to where it is today, one may have to take this individual variability

into account. Social behavior is also transmitted from generation to generation, both biologically and culturally, and therefore also evolves. For instance, there are many examples of plant and animal species in which the biology of the plants and the sociology of the animals clearly coevolved, to the benefit of both. Game-theory models have been used to study the evolution of certain human behaviors such as altruism. Therefore, ecological problems, which sometimes seem simple at first, are often very deep, because the biology and its evolution are connected in complicated ways to both the physics of the environment and the social behavior of the animals. A good introductory review of these questions can be found in Levin et al. (1997).

9 Phylogenetics and Graph Theory

Since Darwin, a deep ongoing problem in biology has been to determine the history of the evolution of species that has brought us to our current state. It is natural when thinking about such questions to draw directed GRAPHS [III.34] in which the vertices, V , are species (past or present) and an edge from species v_1 to species v_2 indicates that v_2 evolved directly from v_1 . Indeed, Darwin himself wrote down such graphs. To explain the mathematical issues, we will consider a simple special case. A connected graph with no cycles is called a *tree*. If we distinguish a particular vertex, ρ , and call it the *root*, then the tree is called *rooted*. The vertices of the tree that have degree one (i.e., have only one attached edge) are called *leaves*. We will assume that ρ is not a leaf. Notice that, because there are no cycles, there is exactly one path in the tree from ρ to each vertex v . We say that $v_1 \leq v_2$ if the path from ρ to v_2 contains v_1 (see figure 2). The problem is to determine which trees with a given set of leaves X (current species) and a given root vertex ρ (a hypothesized ancestral species) are consistent with experimental information and theoretical assumptions about the mechanisms of evolution. Such a tree is called a *rooted phylogenetic X-tree*. One can always add extra intermediate species, so typically one imposes the additional restriction that the phylogenetic trees be as simple as possible.

Suppose that we are interested in a certain characteristic, the number of teeth, for example. We can use it to define a function f from X , the set of current species, to the nonnegative integers: given a species x in X , we let $f(x)$ be the number of teeth of members of x . In general, a *character* is a function from X to a set C of possible values of a particular characteristic (having or not

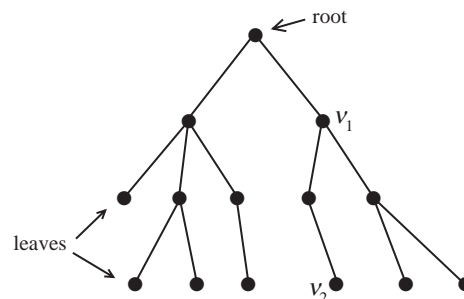


Figure 2 A rooted tree.

having a gene, the number of vertebrae, the presence or absence of a particular enzyme, etc.). It is characters such as these that are measured by biologists in current species. In order to say something about evolutionary history, one would like to extend the definition of f from X to the larger set V of all the vertices in a phylogenetic tree. To do this, one specifies some rules for how characters can change as species evolve. A character is called *convex* if f can be extended to a function \tilde{f} from V to C in such a way that for every $c \in C$, the subset $\tilde{f}^{-1}(c)$ of V is a connected subgraph of the tree. That is, between any two species x and y with character value c there should be a path back in evolutionary history from x and forward again to y such that all the species in between have the same value c . This essentially forbids new values from arising and then reverting back and forbids two values evolving separately (in different parts of the tree). Of course, we have the current species and lots of characters. What is unknown is the phylogenetic tree, that is, the collection of intermediate species and the relations between them that link the current species to a common ancestor. A collection of characters is called *compatible* if there exists a phylogenetic tree on which they are all convex. Determining when this is the case and finding an algorithm for constructing such a tree (or a minimal such tree) is called the *perfect phylogeny problem*. This problem is understood for collections of characters with binary values, but not in general.

An alternative problem is the following. Note that we have been treating all the edges alike when in fact some may represent longer or shorter evolutionary steps. Suppose that we have a function w that assigns a positive number to each edge. Then, since there is a unique shortest path between any two vertices in the tree, w induces a distance function d_w on $V \times V$, and in particular on X . Now, suppose that we are given a distance

function δ on $X \times X$ that tells us how far apart current species are. The question is whether there exists a phylogenetic tree and a weighting function w so that $\delta(x, y) = d_w(x, y)$ for all $x, y \in X$. If so, one would like an algorithm to construct the tree and the weights. If not, one would like to construct a family of trees that satisfy the relation approximately.

Finally, we note that there is a blossoming field of Markov processes on trees where the partial order on V forms the basis for the Markov condition. Not only are there wonderful mathematical questions relating the geometry of the tree to the processes, but there are important issues for phylogenetics. Suppose that one starts with characters defined only at the root and then allows them to “evolve” down the tree by (possibly different) Markov processes. Then, given the distribution of characters on the leaves, when can we reconstruct the tree? These questions have even given rise to problems in algebraic geometry.

Phylogenetics is useful not only for determining our past but also for controlling our present and future: see Fitch et al. (1997), where you can find a phylogenetic reconstruction for the influenza A virus. An excellent recent graduate text in this field is Semple and Steel (2003).

10 Mathematics in Medicine

It is clear that an improved understanding of biological systems leads, at least indirectly, to improved medical care. However, there are many cases in which mathematics has a direct impact on medicine. We give two brief examples.

Charles Taylor is a biomedical engineer at Stanford who works on the fluid dynamics of the cardiovascular system. He wants to use fast simulations of flows as part of the medical decision-making process. Suppose that a patient presents with leg weakness and is found on magnetic resonance imaging (MRI) to have an arterial constriction in the thigh. Typically, the surgical group will meet and consider a variety of options including shunting blood from other vessels to a point below the constriction or shunting blood around the constriction with vessels removed from some other site in the patient’s body. Among a fairly large number of possible choices, the surgical group chooses based on what they have been taught and on their own experience. The characteristics of the flow after the graft are important not just for recovery of function but to prevent the formation of possibly destructive clots. An

important difficulty is that patients treated successfully are rarely seen again, so one does not know the actual characteristics of the flow after the operation. Charles Taylor wants to be in on the discussion with the surgical team with immediate fluid dynamical simulations based on the patient’s actual vasculature (as revealed by the MRI) for each proposed graft suggested. And he wants followup on each patient to check how well his simulations predicted the actual postoperative flow.

David Eddy is an applied mathematician who has worked on health policy for thirty years. He first became prominent when he published *Screening for Cancer: Theory, Analysis and Design* (Eddy 1980), which grew out of his Ph.D. thesis. Because of this book, the American Cancer Society changed its recommendation for the frequency of Pap smears from once a year to once every three years, since Eddy’s modeling showed that the change would have little effect on the life expectancy of the average American woman. A short calculation easily estimates the amount of money saved in an economy that spends 15% of its gross domestic product (GDP) on health care. Throughout his career Eddy has criticized both the indiscriminate use of diagnostic tests and the incorrect use of the results by physicians and policy boards often ignorant of the basic facts of conditional probability. He has criticized specific health-policy guidelines as based on seat-of-the-pants guesswork instead of quantitative analysis. In a classic case he distributed questionnaires to physicians at a conference on colorectal cancer. The physicians were asked to estimate the percentage drop in mortality from colorectal cancers if all Americans over age fifty were to have the two most common diagnostic tests each year: fecal blood smear and flexible sigmoidoscopy. The answers were approximately uniformly distributed in a range from 2% to 95%. Even more startling was the fact that the physicians did not even know that they disagreed so dramatically. He has used mathematical models to analyze the costs and benefits of new and existing surgeries, medical treatments, and drugs, and he has participated robustly in debates on the current health-policy crisis. Throughout, he has pointed out that a hefty percentage of GDP is spent on devices, drugs, and procedures with almost no mathematical analysis of which are effective.

For more on the interrelations between mathematics and medicine, see MATHEMATICS AND MEDICAL STATISTICS [VII.11].

11 Conclusions

Mathematics and mathematicians have played important roles in many fields of biology that this brief article has not had the space to cover: immunology, radiology, developmental biology, and the design of medical devices and synthetic biomaterials, to name just a few of the most obvious omissions. Nevertheless, this collection of examples and introductory discussions allows us to draw a few conclusions about mathematical biology. The range of biological problems needing explanation by mathematics is enormous and techniques from many different branches of mathematics are important. It is not so easy in mathematical biology to extract simple, clear mathematical questions to work on, because biological systems typically operate in a complex environment where it is difficult to decide what should be counted as the system and what as the parts. Finally, biology is a source of new, interesting, and difficult questions for mathematicians, whose participation in the biological revolution is necessary for a full understanding of the biology itself.

Further Reading

- Durrett, R., and S. Levin. 1994. The importance of being discrete (and spatial). *Theoretical Population Biology* 46: 363–94.
- Eddy, D. M. 1980. *Screening for Cancer: Theory, Analysis and Design*. Englewood Cliffs, NJ: Prentice-Hall.
- Fall, C., E. Marland, J. Wagner, and J. Tyson. 2002. *Computational Cell Biology*. New York: Springer.
- Fitch, W. M., R. M. Bush, C. A. Bender, and N. J. Cox. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proceedings of the National Academy of Sciences of the United States of America* 94:7712–18.
- Grossberg, S. 1982. *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition, and Motor Control*. Boston, MA: Kluwer.
- Keener, J., and J. Sneyd. 1998. *Mathematical Physiology*. New York: Springer.
- Levin, S., B. Grenfell, A. Hastings, and A. Perelson. 1997. Mathematical and computational challenges in population biology and ecosystems science. *Science* 275:334–43.
- Nijhout, H. F. 2002. The nature of robustness in development. *Bioessays* 24(6):553–63.
- Pevzner, P. A. 2000. *Computational Molecular Biology: An Algorithmic Approach*. Cambridge, MA: MIT Press.
- Schlick, T. 2002. *Molecular Modeling and Simulation*. New York: Springer.
- Temple, C., and M. Steel. 2003. *Phylogenetics*. Oxford: Oxford University Press.
- Waterman, M. S. 1995. *Introduction to Computational Biology: Maps, Sequences, and Genomes*. London: Chapman and Hall.

VII.3 Wavelets and Applications

Ingrid Daubechies

1 Introduction

One of the best ways to understand a function is to expand it in terms of a well-chosen set of “basic” functions, of which TRIGONOMETRIC FUNCTIONS [III.94] are perhaps the best-known example. Wavelets are families of functions that are very good building blocks for a number of purposes. They emerged in the 1980s from a synthesis of older ideas in mathematics, physics, electrical engineering, and computer science, and have since found applications in a wide range of fields. The following example, concerning image compression, illustrates several important properties of wavelets.

2 Compressing an Image

Directly storing an image on a computer uses a lot of memory. Since memory is a limited resource, it is highly desirable to find more efficient ways of storing images, or rather to find *compressions* of images. One of the main ways of doing this is to express the image as a function and write that function as a linear combination of basic functions of some kind. Typically, most of the coefficients in the expansion will be small, and if the basic functions are chosen in a good way it may well be that one can change all these small coefficients to zero without changing the original function in a way that is visually detectable.

Digital images are typically given by large collections of *pixels* (short for *picture elements*; see figure 1).

The boat image in figure 1 is made up of 256×384 pixels; each pixel has one of 256 possible gray values, ranging from pitch black to pure white. (Similar ideas apply to color images, but for this exposition, it is simpler to keep track of only one color.) Writing a number between 0 and 255 requires 8 digits in binary; the resulting 8-bit requirement to register the gray level for each of the $256 \times 384 = 98\,304$ pixels thus gives a total memory requirement of 786 432 bits, for just this one image.

This memory requirement can be significantly reduced. In figure 2, two squares of 36×36 pixels are highlighted, in different areas of the image. As is clear from its blowup, square A has fewer distinctive characteristics than square B (a blowup of which is shown in figure 1), and should therefore be describable with fewer

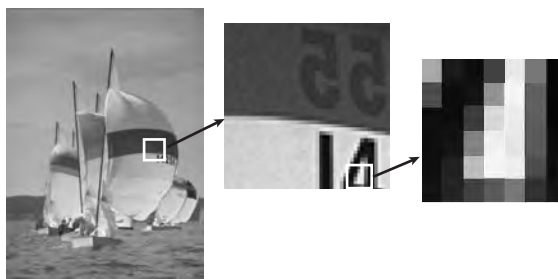


Figure 1 A digital image with successive blowups.

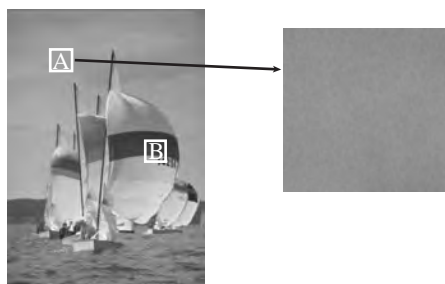


Figure 2 Blowup of a 36×36 square in the sky.

bits. Square B has more features, but it too contains (smaller) squares that consist of many similar pixels; again this can be used to describe this region with fewer than the $36 \times 36 \times 8$ bits given by the naive estimate of assigning 8 bits to each pixel.

These arguments suggest that a change in the representation of the image can lead to reduced memory requirements: instead of a huge assembly of pixels, all equally small, the image should be viewed as a combination of regions of different size, each of which has more or less constant gray value; each such region can then be described by its size (or scale), by where it appears in the image, and by the 8-bit number that tells us its average gray value. Given any subregion of the image, it is easy to check whether it is already of this simple type by comparing it with its average gray value. For square A, taking the average makes virtually no difference, but for square B, the average gray value is not sufficient to characterize this portion of the image (see figure 3).

When square B is subdivided into smaller subsquares, some of them have a virtually constant gray level (e.g., in the top-left or bottom-left regions of square B); others, such as subsquares 2 and 3 (see figure 4), that are not of just one constant gray level may

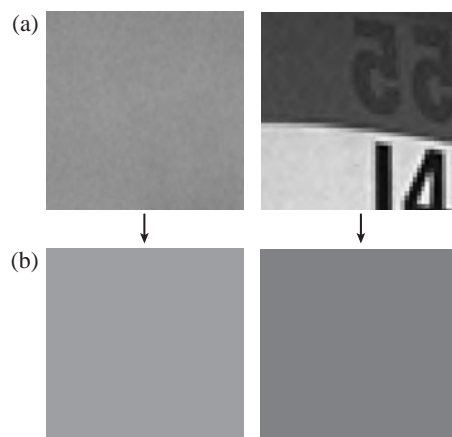


Figure 3 (a) Blowups of squares A (left) and B (right) with (b) the average gray value for each.

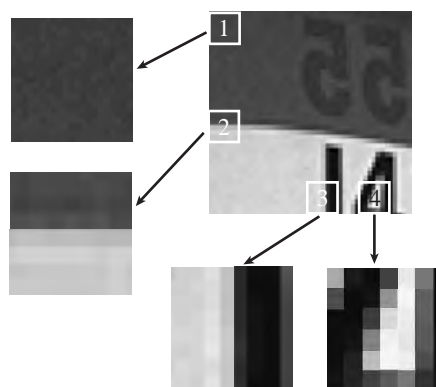


Figure 4 Subsquare 1 has constant gray level, while subsquares 2 and 3 do not, but they can be split horizontally (2) or vertically (3) into two regions with (almost) constant gray level. Subsquare 4 needs finer subdivision to be reduced to “simple” regions.

still have a simple gray level substructure that can be easily characterized with a few bits.

To use this decomposition for image compression, one should be able to implement it easily in an automated way. This could be done as follows:

- first, determine the average gray value for the whole image (assumed to be square, for simplicity);
- compare a square with this constant gray value with the original image; if it is close enough, then we are done (but it will have been a very boring image);

- if more features are needed than only the average gray value, subdivide the image into four equal-sized squares;
- for each of these subsquares, determine *their* average gray value, and compare with the subsquare itself;
- for those subsquares that are not sufficiently characterized by their average gray value, subdivide again into four further equal-sized subsquares (each now having an area one sixteenth of the original image);
- and so on.

In some of the subsquares it may be necessary to divide down to the pixel level (as in subsquare 4 in figure 4, for example), but in most cases subdivision can be stopped much earlier. Although this method is very easy to implement automatically, and leads to a description using many fewer bits for images such as the one shown, it is still somewhat wasteful. For instance, if the average gray level of the original image is 160, and we next determine the gray levels of each of the four quarter images as 224, 176, 112, and 128, then we have really computed one number too many: the average of the gray levels for the four equal-sized subimages is automatically the gray level of the whole image, so it is unnecessary to store all five numbers. In addition to the average gray value for a square, one just needs to store the *extra* information contained in the average gray values of its four quarters, given by the three numbers that describe

- how much darker (or lighter) the left half of the square is than the right,
- how much darker (or lighter) the top half of the square is than the bottom, and
- how much darker (or lighter) the diagonal from lower left to upper right is than the other diagonal.

Consider for example a square divided up into four subsquares with average values 224, 176, 112, and 128, as shown in figure 5. The average gray value for the whole square can easily be checked to be 160. Now let us do three further calculations. First, we work out the average gray values of the top half and the bottom half, which are 200 and 120, respectively, and calculate their difference, which is 80. Then we do the same for the left half and the right half, obtaining the difference $168 - 152 = 16$. Finally, we divide the four squares up diagonally: the average over the bottom-left and top-

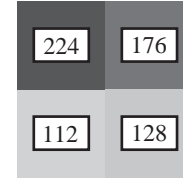


Figure 5 The average gray values for four subsquares of a square.

right squares is 144, the average over the other two is 176, and the difference between these two is -32 .

From these four numbers one can reconstruct the four original averages. For example, the average for the top-right subsquare is given by $160 + [80 - 16 + (-32)]/2 = 176$.

It is thus *this* process, rather than simply averaging over smaller and smaller squares as described above, that needs to be repeated. We now turn to the question of making the whole decomposition procedure as efficient as possible.

A complete decomposition of a 256×256 square, from “top” (largest square) to “bottom” (the three types of “differences” for the 2×2 subsquares), involves the computation of many numbers (in fact exactly 256×256 before pruning), some of which are themselves combinations of many of the original pixel values. For instance, the grayscale average of the whole 256×256 square requires adding $256 \times 256 = 65\,536$ numbers with values between 0 and 255 and then dividing the result by 65 536; another example, the difference between the averages of the left and right halves, requires adding the $256 \times 128 = 32\,768$ grayscale numbers for the left half and then subtracting from this sum A the sum B of another 32 768 numbers. On the other hand, the sum of the pixel grayscale values over the whole square is simply $A + B$, a sum of two 33-bit numbers instead of 65 536 numbers of 8 bits each. This allows us to make a considerable saving in computational complexity if A and B are computed *before* the average over the whole square. A computationally optimal implementation of the ideas explained so far must therefore proceed along a different path from the one sketched above.

Indeed, a much better procedure is to start from the other end of the scale. Instead of starting with the whole image and repeatedly subdividing it, one begins at the pixel level and builds up. If the image has $2^J \times 2^J$ pixels in total, then it can also be viewed as consisting of $2^{J-1} \times 2^{J-1}$ “superpixels,” each of which is a small

square of 2×2 pixels. For each 2×2 square, the average of the four gray values can be computed (this is the gray value of the superpixel), as well as the three types of differences indicated above. Moreover, these computations are all very simple.

The next step is to store the three difference values for each of the 2×2 squares and organize their averages, the gray values of the $2^{J-1} \times 2^{J-1}$ superpixels, into a new square. This square can be divided, in turn, into $2^{J-2} \times 2^{J-2}$ “super-superpixels,” each of which is a small square of 2×2 superpixels (and thus stands for 4×4 “standard” pixels), and so on. At the very end, after J levels of “zooming out,” there is only one super J -pixel remaining; its gray value is the average over the whole image. The *last* three differences that were computed in this pixel-level-up process correspond exactly to the largest-scale differences that the top-down procedure would have computed *first*, at much greater computational expense.

Carrying out the procedure from the pixel level up, none of the individual averaging or differencing computations involves more than two numbers; the total number of these elementary computations, for the *whole* transform, is only $8(2^{2J} - 1)/3$. For the 256×256 square discussed before, $J = 8$, so the total is 174 752, which is about the number of computations needed for just one level in the top-down procedure.

How can all this lead to compression? At each stage of the process, three species of difference numbers are accumulated, at different levels and corresponding to different positions. The total number of differences calculated is $3(1 + 2^2 + \dots + 2^{2(J-1)}) = 2^{2J} - 1$. Together with the gray value of the whole square, this means we end up with exactly as many numbers as we had gray values for the original $2^J \times 2^J$ pixels. However, many of these difference numbers will be very small (as argued before), and can just as well be dropped or put to zero, and if the image is reconstructed from the remainder there will be no perceptible loss of quality. Once we have set these very small differences to zero, a list that enumerates all the differences (in some prearranged order) can be made much shorter: whenever a long stretch of Z zeros is encountered, it can be replaced by the statement “insert Z zeros now,” which requires only a prearranged symbol (for “insert zeros now”), followed by the number of bits needed for Z , i.e., $\log_2 Z$. This achieves, as desired, a significant compression of the data that need to be stored for large images. (In practice, however, image compression involves *many* more issues, to which we shall return briefly below.)

The very simple image decomposition described above is an elementary example of a *wavelet decomposition*. The data that are retained consist of

- a very coarse approximation, and
- additional layers giving detail at successively finer scales j , with j ranging from 0 (the coarsest level) to $J - 1$ (the first superpixel level).

Moreover, within each scale j the detail layer consists of many pieces, each of which has a definite localization (indicating to which of the super j -pixels it pertains), and all the pieces have “size” 2^j . (That is, the size, in pixel widths, of the corresponding super j -pixel is 2^j .) In particular, the building blocks are very small at fine scales and become gradually larger as the scale becomes coarser.

3 Wavelet Transforms of Functions

In the image-compression example we needed to look at three types of differences at each level (horizontal, vertical, and diagonal) because the example was a two-dimensional image. For a one-dimensional signal, one type of difference suffices. Given a function f from \mathbb{R} to \mathbb{R} , one can write a wavelet transform of f that is entirely analogous to the image example. For simplicity, let us look at a function f such that $f(x) = 0$ except when x belongs to the interval $[0, 1]$.

Let us now consider successive approximations of f by *step functions*: that is, functions that change value in only finitely many places. More precisely, for each positive integer j , divide the interval $[0, 1]$ up into 2^j equal intervals, denoting the interval from $k2^{-j}$ to $(k+1)2^{-j}$ by $I_{j,k}$ (so that k runs from 0 to $2^j - 1$). Then define a function $P_j(f)$ by setting its value on $I_{j,k}$ to be the average value of f on that interval. This is illustrated in figure 6, which shows the step function $P_3(f)$ for a function f whose graph is shown as well. As j increases, the width of the intervals $I_{j,k}$ decreases, and $P_j(f)$ gets closer to f . (In more precise mathematical terms, if $p < \infty$ and f belongs to the FUNCTION SPACE [III.29] L_p , then $P_j(f)$ converges to f in L_p .)

Each approximation $P_j(f)$ of f can be computed easily from the approximation $P_{j+1}(f)$ at the next-finer scale: the average of the values that $P_{j+1}(f)$ takes on the two intervals $I_{j+1,2k}$ and $I_{j+1,2k+1}$ gives the value that $P_j(f)$ takes on $I_{j,k}$.

Of course, some information about f is lost when we move from $P_{j+1}(f)$ to $P_j(f)$. On every interval $I_{j,k}$, the

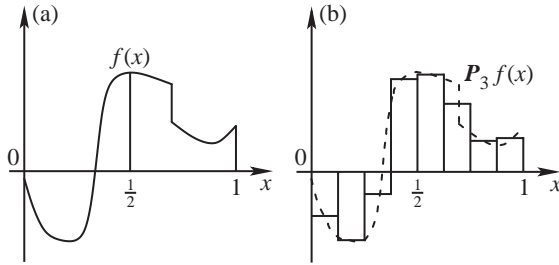


Figure 6 Graphs of (a) the function f and (b) its approximation $P_3(f)$, which is constant on every interval between $l/8$ and $(l+1)/8$, with $l = 0, 1, \dots, 7$, and exactly equal to the average of f on each of these intervals.

difference between $P_{j+1}(f)$ and $P_j(f)$ is a step function, with constant levels on the $I_{j+1,l}$, that takes on exactly opposite values on each pair $(I_{j+1,2k}, I_{j+1,2k+1})$. The difference $P_{j+1}(f) - P_j(f)$ of the two approximation functions, over all of $[0, 1]$, consists of a juxtaposition of such up-and-down (or down-and-up) step functions, and can therefore be written as a sum of translates of the same up-and-down function, with appropriate coefficients:

$$P_{j+1}(f)(x) - P_j(f)(x) = \sum_{k=0}^{2^j-1} a_{j,k} U_j(x - 2^{-j}k),$$

where

$$U_j(x) = \begin{cases} 1 & \text{for } x \text{ between } 0 \text{ and } 2^{-(j+1)}, \\ -1 & \text{for } x \text{ between } 2^{-(j+1)} \text{ and } 2 \times 2^{-(j+1)}, \\ 0 & \text{for all other } x. \end{cases}$$

Moreover, the “difference functions” U_j at the different levels are all scaled copies of a single function H , which takes the value -1 between 0 and $\frac{1}{2}$ and $+1$ between $\frac{1}{2}$ and 1 ; indeed, $U_j(x) = H(2^j x)$. It follows that each difference $P_{j+1}(f)(x) - P_j(f)(x)$ is a linear combination of the functions $H(2^j x - k)$, with k ranging from 0 to $2^j - 1$; adding many such differences, for successive j , shows that $P_J(f)(x) - P_0(f)(x)$ is a linear combination of the collection of functions $H(2^j x - k)$, with j ranging from 0 to $J - 1$ and k ranging from 0 to $2^j - 1$. Picking larger and larger J makes $P_J(f)$ closer and closer to f ; one finds that $f - P_0(f)$ (i.e., the difference between f and its average) can be viewed as a (possibly infinite) linear combination of the functions $H(2^j x - k)$, now with j ranging over all the nonnegative integers.

This decomposition is very similar to what was done for images at the start of the article, but in one dimension instead of two and presented in a more abstract

way. The basic ingredients are that f minus its average has been decomposed into a sum of layers at successively finer and finer scales, and that each extra layer of detail consists of a sum of simple “difference contributions” that all have width proportional to the scale. Moreover, this decomposition is realized by using translates and dilates of the *single* function $H(x)$, often called the *Haar wavelet*, after Alfred Haar, who first defined it at the start of the twentieth century (though not in a wavelet context). The functions $H(2^j x - k)$ constitute an *orthogonal* set of functions, meaning that the inner product $\int H(2^j x - k)H(2^{j'} x - k') dx$ is zero except when $j = j'$ and $k = k'$; if we define $H_{j,k}(x) = 2^{j/2}H(2^j x - k)$, then we also have that $\int [H_{j,k}(x)]^2 dx = 1$. A consequence of this is that the *wavelet coefficients* $w_{j,k}(f)$ that appear when we write the “ j th layer” $P_{j+1}(f)(x) - P_j(f)(x)$ of the function f as a linear combination $\sum_k w_{j,k}(f)H_{j,k}(x)$ are given by the formula $w_{j,k}(f) = \int f(x)H_{j,k}(x) dx$.

Haar wavelets are a good tool for exposition, but for most applications, including image compression, they are not the best choice. Basically, this is because replacing a function simply by its averages over intervals (in one dimension) or squares (in two dimensions) results in a very-low-quality approximation, as illustrated in figure 7(b).

As the scale of approximation is made finer and finer (i.e., as the j in $P_j(f)$ increases), the difference between f and $P_j(f)$ becomes smaller; with a piecewise-constant approximation, however, this requires corrections at almost every scale “to get it right” in the end. Unless the original happens to be made up of large areas where it is roughly constant, many small-scale Haar wavelets will be required even in stretches where the function just has a consistent, sustained slope, without “genuine” fine features.

The right framework to discuss these questions is that of *approximation schemes*. An approximation scheme can be defined by providing a family of “building blocks,” often with a natural order in which they are usually enumerated. A common way of measuring the quality of an approximation scheme is to define V_N to be the space of all linear combinations of the first N building blocks, and then to let $A_N f$ be the closest function in V_N to f , where distance is measured by the L_2 -norm (though other norms can also be used). Then one examines how the distance $\|f - A_N f\|_2 = [\int |f(x) - A_N f(x)|^2 dx]^{1/2}$ decays as N tends to infinity. An approximation scheme is said to be of *order* L for a class of functions \mathcal{F} if $\|f - A_N f\|_2 \leq CN^{-L}$ for all

T&T note: check position of this figure at page make-up stage. Note for Terri: ‘ $k2^{-3}$, $(k+1)2^{-3}$ ’ is OK in caption – bracket type has to do with whether the range is inclusive or exclusive. Another note for Terri: change from ‘(b) and (c)’ to ‘(b), (c)’ OK in caption?

Terri: Tim has checked through the notation in this article and thinks it’s all OK and consistent – and has previously checked with the author as he was also uncertain initially.

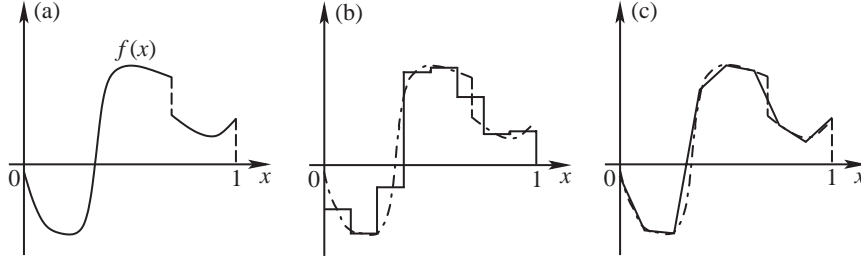


Figure 7 (a) The original function. (b), (c) Approximations of f by a function that equals a polynomial on each interval $[k2^{-3}, (k+1)2^{-3})$. The best approximation of f by a piecewise-constant function is shown in (b); the best by a continuous piecewise-linear function is in (c).

functions f in \mathcal{F} , where C typically depends on f but must be independent of N . The order of an approximation scheme for smooth functions is closely linked to the performance of the approximation scheme on polynomials (because smooth functions can be replaced in estimations, at very little cost, by the polynomials given by their Taylor expansions). In particular, the types of approximation schemes considered here can have order L only if they perfectly reproduce polynomials of degree at most $L - 1$. In other words, there should exist some N_0 such that if p is any polynomial of degree at most $L - 1$ and $N \geq N_0$, then $A_N p = p$.

In the Haar case, applied to functions f that differ from zero only between 0 and 1, the building blocks consist of the function φ that takes the value 1 on $[0, 1]$ and 0 outside, together with the families $\{H_{j,k}; k = 0, \dots, 2^j - 1\}$ for $j = 0, 1, 2, \dots$. We saw above that $P_j^{\text{Haar}}(f)$ can be written as a linear combination of the first $1 + 2^0 + 2^1 + \dots + 2^{j-1} = 2^j$ building blocks $\varphi, H_{0,0}, H_{1,0}, H_{1,1}, H_{2,0}, \dots, H_{j-1,2^{j-1}-1}$. Because the Haar wavelets are orthogonal to each other, this is also the linear combination of these basis functions that is closest to f , so that $P_j^{\text{Haar}}(f) = A_{2^j}^{\text{Haar}} f$. Figure 7 shows (for $j = 3$) both $A_{2^j}^{\text{Haar}} f$ and $A_{2^j}^{\text{PL}} f$, which is the best approximation of f by a continuous, piecewise-linear function with breakpoints at $k2^{-j}, k = 0, 1, \dots, 2^j - 1$. It turns out that if you are trying to approximate a function f using Haar wavelets, then the best decay you can obtain, even if f is smooth, is of the form $\|f - P_j^{\text{Haar}}(f)\|_2 \leq C2^{-j}$, or $\|f - A_N^{\text{Haar}} f\|_2 \leq CN^{-1}$ for $N = 2^j$. This means that approximation by Haar wavelets is a *first-order* approximation scheme. Approximation by continuous piecewise-linear functions is a *second-order* scheme: for smooth f , $\|f - A_N^{\text{PL}} f\|_2 \leq CN^{-2}$ for $N = 2^j$. Note that the difference between the two schemes can also be seen from the maximal degree d of polynomials they “reproduce” perfectly: clearly both schemes can repro-

duce constants ($d = 0$); the piecewise-linear scheme can also reproduce linear functions ($d = 1$), whereas the Haar scheme cannot.

Take now any continuously differentiable function f defined on the interval $[0, 1]$. Typically $\|f - P_j^{\text{Haar}}(f)\|_2$ equals about $C2^{-j}$; for an approximation scheme of order 2, that same difference would be about $C'2^{-2j}$. In order to achieve the same accuracy as $P_j^{\text{Haar}}(f)$, the piecewise-linear scheme would thus require only $j/2$ levels instead of j levels. For higher orders L , the gain would be even greater. If the projections P_j gave rise to a higher-order approximation scheme like this, then the difference $P_{j+1}(f)(x) - P_j(f)(x)$ would be so small as not to matter, even for modest values of j , wherever the function f was reasonably smooth; for these values of j , the difference would be important only near points where the function was not as smooth, and so only in those places would a contribution be needed from “difference coefficients” at very fine scales.

This is a powerful motivation to develop a framework similar to that for Haar, but with fancier “generalized averages and differences” corresponding to successive $P_j(f)$ associated with higher-order approximation schemes. This can be done, and was done in an exciting period in the 1980s to which we shall return briefly below. In these constructions, the generalized averages and differences are typically computed by combining more than two finer-scale entries each time, in appropriate linear combinations. The corresponding function decomposition represents functions as (possibly infinite) linear combinations of wavelets $\psi_{j,k}$ derived from a wavelet ψ . As in the case of H , $\psi_{j,k}(x)$ is defined to be $2^{j/2}\psi(2^j x - k)$. Thus, the functions $\psi_{j,k}$ are again normalized translates and dilates of a *single* function; this is due to our using systematically the same averaging operator to go from scale $j + 1$ to scale j , and the same differencing operator to quantify

the difference between levels $j + 1$ and j , regardless of the value of j . There is no absolutely compelling reason to use the same averaging and differencing operator for the transition between any two successive levels, and thus to have all the $\psi_{j,k}$ generated by translating and dilating a single function. However, it is very convenient for implementing the transform, and it simplifies the mathematical analysis.

One can additionally require that, like the $H_{j,k}$, the $\psi_{j,k}$ constitute an *orthonormal basis* for the space $L^2(\mathbb{R})$. The *basis* part means that every function can be written as a (possibly infinite) linear combination of the $\psi_{j,k}$; the *orthonormality* means that the $\psi_{j,k}$ are *orthogonal* to each other, except if they are equal, in which case their inner product is 1.

As discussed above, the projections P_j for the wavelet ψ will correspond to an approximation scheme of order L only if they can reproduce perfectly all polynomials of degree less than L . If the functions $\psi_{j,k}$ are orthogonal, then $\int \psi_{j',k}(x) P_j(f)(x) dx = 0$ whenever $j' > j$. The $\psi_{j,k}$ can thus be associated with an approximation scheme of order L only if $\int \psi_{j,k}(x) p(x) dx = 0$ for sufficiently large j and for all polynomials p of degree less than L . By scaling and translating, this reduces to the requirement $\int x^l \psi(x) dx = 0$ for $l = 0, 1, \dots, L - 1$. When this requirement is met, ψ is said to *have L vanishing moments*.

Figure 8 shows the graphs of some choices for ψ that give rise to orthonormal wavelet bases and that are used in various circumstances.

For the wavelets of the type $\psi^{[2n]}$, and thus in particular for $\psi^{[4]}$, $\psi^{[6]}$, and $\psi^{[12]}$ in figure 8, an algorithm similar to that for the Haar wavelet can be used to carry out the decomposition, except that instead of combining two numbers from $P_{j+1,k}$ to obtain an average or a difference coefficient at level j , these wavelet decompositions require weighted combinations of four, six, or twelve finer-level numbers, respectively. (More generally, $2n$ finer-level numbers are used for $\psi^{[2n]}$.)

Because the Meyer and Battle-Lemarié wavelets $\psi^{[M]}$ and $\psi^{[BL]}$ are not concentrated on a finite interval, different algorithms are used for wavelet expansions with respect to these wavelets.

There are many useful orthonormal wavelet bases besides the examples given above. Which one to choose depends on the application one has in mind. For instance, if the function classes of interest in the application have smooth pieces, with abrupt transitions or spikes, then it is advantageous to pick a smooth ψ , corresponding to a high-order approximation scheme.

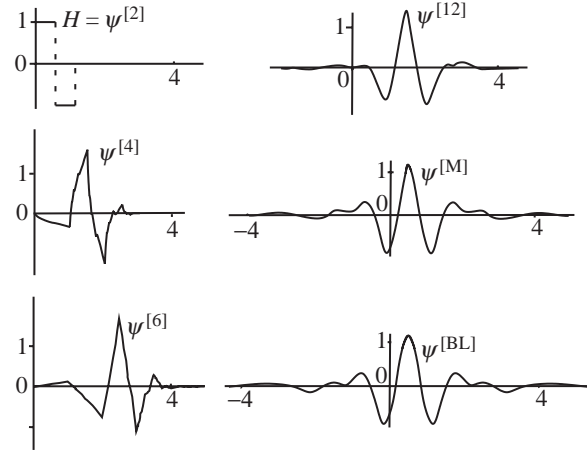


Figure 8 Six different choices of ψ for which the $\psi_{j,k}(x) = 2^{j/2} \psi(2^j x - k)$, $j, k \in \mathbb{Z}$, constitute an orthonormal basis for $L^2(\mathbb{R})$. The Haar wavelet can be viewed as the first example of a family $\psi^{[2n]}$, of which the wavelets for $n = 2, 3$, and 6 are also plotted here. Each $\psi^{[2n]}$ has n vanishing moments and is supported on (i.e., is equal to zero outside) an interval of width $2n - 1$. The remaining two wavelets are not supported on an interval; however, the Fourier transform of the Meyer wavelet $\psi^{[M]}$ is supported on $[-8\pi/3, -2\pi/3] \cup [2\pi/3, 8\pi/3]$; all moments of $\psi^{[M]}$ vanish. The Battle-Lemarié wavelet $\psi^{[BL]}$ is twice differentiable, is piecewise polynomial of degree 3, and has exponential decay; it has four vanishing moments.

This allows one to describe the smooth pieces efficiently with coarse-scale basis functions, and to leave the fine-scale wavelets to deal with the spikes and abrupt transitions. In that case, why not always use a wavelet basis with a very high approximation order? The reason is that most applications require numerical computation of wavelet transforms; the higher the order of the approximation scheme, the more spread out the wavelet, and the more terms have to be used in each generalized average/difference, which slows down numerical computation. In addition, the wider the wavelet, and hence the wider all the finer-scale wavelets derived from it, the more often a discontinuity or sharp transition will overlap with these wavelets. This tends to spread out the influence of such transitions over more fine-scale wavelet coefficients. Therefore, one must find a good balance between the approximation order and the width of the wavelet, and the best balance varies from problem to problem.

There are also wavelet bases in which the restriction of orthonormality is relaxed. In this case one typically uses two different “dual” wavelets ψ and $\tilde{\psi}$, such that

T&T note: must check forced linebreak in caption for this figure immediately prior to press stage.

$\int_{-\infty}^{\infty} \psi_{j,k}(x) \tilde{\psi}_{j',k'}(x) dx = 0$ unless $j = j'$ and $k = k'$. The approximation order of the scheme that approximates functions f by linear combinations of the $\psi_{j,k}$ is then governed by the number of vanishing moments of $\tilde{\psi}$. Such wavelet bases are called *biorthogonal*. They have the advantage that the basic wavelets ψ and $\tilde{\psi}$ can both be symmetric and concentrated on an interval, which is impossible for orthonormal wavelet bases other than the Haar wavelets.

The symmetry condition is important for image decomposition, where preference is usually given to two-dimensional wavelet bases derived from one-dimensional bases with a symmetric function ψ , a derivation to which we return below. When an image is compressed by deleting or rounding off wavelet coefficients, the difference between the original image I and its compressed version I^{comp} is a combination, with small coefficients, of these two-dimensional wavelets. It has been observed that the human visual system is more tolerant of such small deviations if they are symmetric; the use of symmetric wavelets thus allows for slightly larger errors, which translates to higher compression rates, before the deviations cross the threshold of perception or acceptability.

Another way of generalizing the notion of wavelet bases is to allow more than one starting wavelet. Such systems, known as *multiwavelets*, can be useful even in one dimension.

When wavelet bases are considered for functions defined on the interval $[a, b]$ rather than the whole of \mathbb{R} , the constructions are typically adapted, giving bases of *interval wavelets* in which specially crafted wavelets are used near the edges of the interval. It is sometimes useful to choose less regular ways of subdividing intervals than the systematic halving considered above: in this case, the constructions can be adapted to give *irregularly spaced* wavelet bases.

When the goal of a decomposition is compression of the information, as in the image example at the start, it is best to use a decomposition that is itself as efficient as possible. For other applications, such as pattern recognition, it is often better to use *redundant* families of wavelets, i.e., collections of wavelets that contain “too many” wavelets, in the sense that all functions in $L^2(\mathbb{R})$ could still be represented even if one dropped some of the wavelets from the collection. *Continuous wavelet families* and *wavelet frames* are the two main kinds of collections used for such redundant wavelet representations.

4 Wavelets and Function Properties

Wavelet expansions are useful for image compression because many regions of an image do not have features at very fine scales. Returning to the one-dimensional case, the same is true for a function that is reasonably smooth at most but not all points, like the function illustrated in figure 6(a). If we zoom in on such a function near a point x_0 where it is smooth, then it will look almost linear, so we will be able to represent that part of the function efficiently if our wavelets are good at representing linear functions.

This is where wavelet bases other than Haar show their power: the wavelets $\psi^{[4]}$, $\psi^{[6]}$, $\psi^{[12]}$, $\psi^{[M]}$, and $\psi^{[BL]}$ shown in figure 8 all define approximation schemes of order 2 or higher, so that $\int x \psi_{j,k}(x) dx = 0$ for all j, k . This is also seen in the numerical implementation schemes: the corresponding generalized differencing that computes the wavelet coefficients of f gives a zero result not only when the graph is flat, but also when it is a straight but sloped line, which is not true for the simple differencing used for the Haar basis. As a result, the number of coefficients needed for the wavelet expansion of smooth functions f to reach a preassigned accuracy is much smaller when one uses more sophisticated wavelets than the Haar wavelets.

For a function f that is twice differentiable except at a finite number of discontinuities, and with a basic wavelet that has, say, three vanishing moments, typically only very few wavelets at fine scales will be needed to write a very-high-precision approximation to f . Moreover, those will be needed only near the discontinuity points. This feature is characteristic for all wavelet expansions, whether they are with respect to an orthonormal basis, a basis that is nonorthogonal, or even a redundant family.

Figure 9 illustrates this for one type of redundant expansion, which uses the so-called *Mexican hat wavelets*, which are given by $\psi(x) = (2\sqrt{2}/\sqrt{3})\pi^{-1/4}(1 - 4x^2)e^{-2x^2}$; this wavelet gets its name from the shape of its graph, which looks like the cross section of a Mexican hat (see the figure).

The smoother a function f is (i.e., the more times it is differentiable), the faster its wavelet coefficients will decay as j increases, provided the wavelet ψ has sufficiently many vanishing moments. The converse statement is also true: one can read off how smooth the function is at x_0 from how the wavelet coefficients $w_{j,k}(f)$ decay, as j increases. Here one restricts attention to the “relevant” pairs (j, k) . In other words, one considers

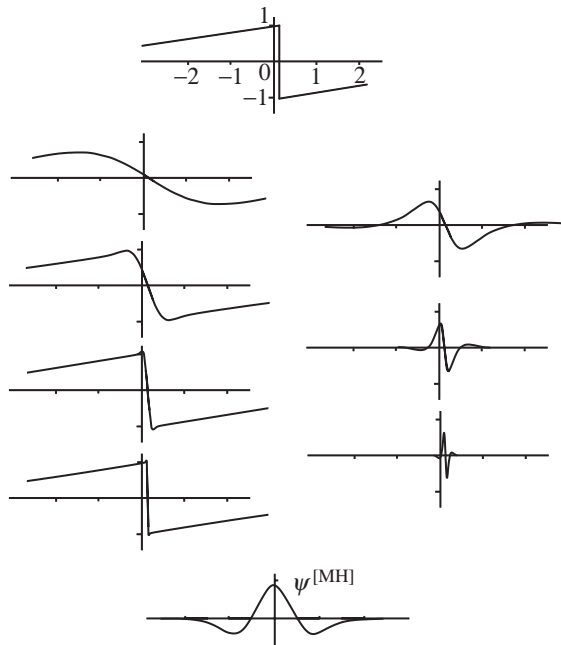


Figure 9 A function with a single discontinuity (top) is approximated by finite linear combinations of Mexican hat wavelets $\psi_{j,l}^{[MH]}$; the graph of $\psi^{[MH]}$ is at the bottom of the figure. Adding finer scales leads to increased precision. Left: successive approximations for $j = 1, 3, 5$, and 7 . Right: total contributions from the wavelets at the scales needed to bridge from one j to the next. (In this example, j increases in steps of $\frac{1}{2}$.) The finer the scale, the more the extra detail is concentrated near the discontinuity point.

only the pairs where $\psi_{j,k}$ is localized near x_0 . (In more precise terms, this converse statement can be reformulated as an exact characterization of the so-called *Lipschitz spaces* C^α , for all noninteger α that are strictly less than the number of vanishing moments of ψ .)

Wavelet coefficients can be used to characterize many other useful properties of functions, both global and local. Because of this, wavelets are good bases not just for L^2 -spaces or the Lipschitz spaces, but also for many other function spaces, such as, for instance, the L^p -spaces with $1 < p < \infty$, the SOBOLEV SPACES [III.29 §2.4], and a wide range of Besov spaces. The versatility of wavelets is partly due to their connection with powerful techniques developed in harmonic analysis throughout the twentieth century.

We have already seen in some detail that wavelet bases are associated with approximation schemes of different orders. So far we have considered approximation schemes in which the $A_N f$ are always linear

combinations of the same N building blocks, regardless of the function f . This is called *linear* approximation, because the collection of all functions of the form $A_N f$ is contained in the linear span V_N of the first N basis functions. Some of the function spaces mentioned above can be characterized by specifying the decay of $\|f - A_N f\|_2$ as N increases, where A_N is defined in terms of an appropriate wavelet basis.

However, when it is compression that we are interested in, we are really carrying out a different kind of approximation. Given a function f , and a desired accuracy, we want to approximate f to within that accuracy by a linear combination of as few basis functions as possible, but we are not trying to choose those functions from the first few levels. In other words, we are no longer interested in the ordering of the basis functions and we do not prefer one label (j, k) over another.

If we want to formalize this, we can define an approximation $\mathcal{A}_N f$ to be the closest linear combination to f that is made up of at most N basis functions. By analogy with linear approximation, we can then define the set \mathcal{V}_N as the set of all possible linear combinations of N basis functions. However, the sets \mathcal{V}_N are no longer linear spaces: two arbitrary elements \mathcal{V}_N are typically combinations of two different collections of N basis functions, so that their sum has no reason to belong to \mathcal{V}_N (though it will belong to \mathcal{V}_{2N}). For this reason, $\mathcal{A}_N f$ is called a *nonlinear* approximation of f .

One can go further and define classes of functions by imposing conditions on the decay of $\|f - \mathcal{A}_N f\|$, as N increases, with respect to some function space norm $\|\cdot\|$. This can of course be done starting from any basis; wavelet bases distinguish themselves from many other bases (such as the trigonometric functions) in that the resulting function spaces turn out to be standard function spaces, such as the Besov spaces, for example. We have referred several times to functions that are smooth in many places but have possible discontinuities in isolated points, and argued that they can be approximated well by linear combinations of a fairly small number of wavelets. Such functions are special cases of elements of particular Besov spaces, and their good approximation properties by sparse wavelet expansions can be viewed as a consequence of the characterization of these Besov spaces by nonlinear approximation schemes using wavelets.¹

Terri: Tim thinks this sentence is fine. OK?

1. More types of wavelet families, as well as many generalizations, can be found on the Internet at www.wavelet.org.

5 Wavelets in More than One Dimension

There are many ways to extend the one-dimensional constructions to higher dimensions. An easy way to construct a multidimensional wavelet basis is to combine several one-dimensional wavelet bases. The image decomposition at the start is an example of such a combination: it combines two one-dimensional Haar decompositions. We saw earlier that a 2×2 superpixel could be decomposed as follows. First, think of it as arranged in two rows of two numbers, representing the gray levels of the corresponding pixels. Next, for each row replace its two numbers by their average and their difference, obtaining a new 2×2 array. Finally, do the same process to the columns of the new array. This produces four numbers, the result of, respectively,

- averaging both horizontally and vertically,
- averaging horizontally and differencing vertically,
- differencing horizontally and averaging vertically, and
- differencing both horizontally and vertically.

The first is the average gray level for the superpixel, which is needed as the input for the next round of the decomposition at the next scale up. The other three correspond to the three types of “differences” already encountered earlier. If we start with a rectangular image that consists of 2^K rows, each containing 2^J pixels, then we end up with $2^{K-1} \times 2^{J-1}$ numbers of each of the four types. Each collection is naturally arranged in a rectangle of half the size of the original (in both directions); it is customary in the image-processing literature to put the rectangle with gray values for the superpixels in the top left; the other three rectangles each group together all the differences (or wavelet coefficients) of the other three kinds. (See the level 1 decomposition in figure 10.) The rectangle that results from horizontal differencing and vertical averaging typically has large coefficients at places where the original image has vertical edges (such as the boat masts in the example above); likewise, the horizontal averaging/vertical differencing rectangle has large coefficients for horizontal edges in the original (such as the stripes in the sails); the horizontal differencing/vertical differencing rectangle selects for diagonal features. The three different types of “difference terms” indicate that we have here three basic wavelets (instead of just one in the one-dimensional case).

In order to go to the next round, one scale up, the scenario is repeated on the rectangle that contains the

superpixel gray values (the results of averaging both horizontally and vertically); the other three rectangles are left unchanged. Figure 10 shows the result of this process for the original boat image, though the wavelet basis used here is not the Haar basis, but a symmetric biorthogonal wavelet basis that has been adopted in the JPEG 2000 image compression standard. The result is a decomposition of the original image into its component wavelets. The fact that so much of this is gray indicates that a lot of this information can be discarded without affecting the image quality.

Figure 11 illustrates that the number of vanishing moments is important not just when the wavelet basis is used for characterizing properties of functions, but also when it comes to image analysis. It shows an image that has been decomposed in two different ways: once with Haar wavelets, the other with the JPEG 2000 standard biorthogonal wavelet basis. In both cases, all but the largest 5% of the wavelet coefficients have been set to zero, and we are looking at the corresponding reconstructions of the images, neither of which is perfect. However, the wavelet used in the JPEG 2000 standard has four vanishing moments, and therefore gives a much better approximation in smoothly varying parts of the image than the Haar basis. Moreover, the reconstruction obtained from the Haar expansion is “blockier” and less attractive.

6 Truth in Advertising: Closer to True Image Compression

Image compression has been discussed several times in this article, and it is indeed a context in which wavelets are used. However, in practice there is much more to image compression than the simple idea of dropping all but the largest wavelet coefficients, taking the resulting truncated list of coefficients, and replacing each of the many long stretches of zeros by its runlength. In this short section we shall give a glimpse of the large gap between the mathematical theory of wavelets as discussed above and the real-life practice of engineers who want to compress images.

First of all, compression applications set a “bit budget,” and all the information to be stored has to fit within the bit budget; statistical estimates and information-theoretic arguments about the class of images under consideration are used to allocate different numbers of bits to different types of coefficients. This bit allocation is much more gradual and subtle than just retaining or dropping coefficients. Even so,

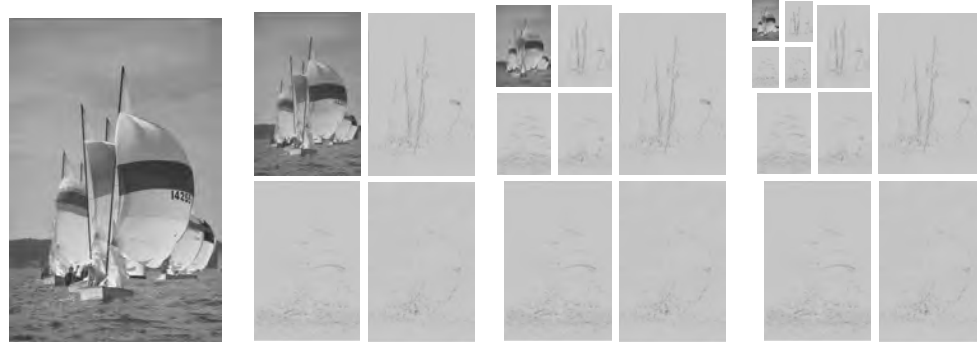


Figure 10 Wavelet decomposition of the boat image, together with a grayscale rendition of the wavelet coefficients. The decompositions are shown after one level of averaging and differencing, as well as after two and three levels. In the rectangles corresponding to wavelet coefficients (i.e., not averaged in both directions), where numbers can be negative, the convention is to use gray scale 128 for zero, and darker/lighter gray scales for positive and negative values. The wavelet rectangles are mostly at gray scale 128, indicating that most of the wavelet coefficients are negligibly small.

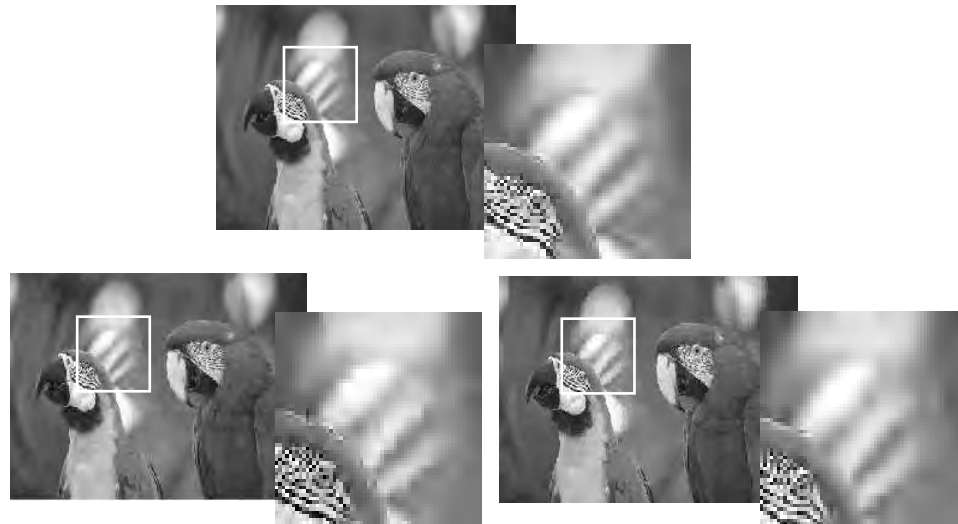


Figure 11 Top: original image, with blowup. Bottom: approximations obtained by expanding the image into a wavelet basis, and discarding the 95% smallest wavelet coefficients. Left: Haar wavelet transform. Right: wavelet transform using the so-called 9-7 biorthogonal wavelet basis.

many coefficients will get no bits assigned to them, meaning that they are indeed dropped altogether.

Because some coefficients are dropped, care has to be taken that each of the remaining coefficients is given its correct *address*, i.e., its (j, k_1, k_2) label, which is essential for “decompressing” the stored information in order to reconstruct the image (or rather, an approximation to it). If you do not have a good strategy for doing this, then you can easily find that the computational resources needed to encode informa-

tion about the addresses cancel out a large portion of the gain made by the nonlinear wavelet approximation. Every practical wavelet-based image-compression scheme uses some sort of clever approach to deal with this problem. One implementation exploits the observation that at locations in the image where wavelet coefficients of some species are negligibly small at some scale j , the wavelet coefficients of the same species at finer scales are often very small as well. (Check it out on the boat image decomposition given above.) At each

such location, this method sets a whole tree of finer-scale coefficients (four for scale $j + 1$, sixteen for scale $j + 2$, etc.) automatically to zero; for those locations where this assumption is not borne out by the wavelet coefficients that are obtained from the actual decomposition of the image at hand, extra bits must then be spent to store the information that a correction has to be made to the assumption. In practice, the bits gained by the “zero-trees” far outweigh the bits needed for these occasional corrections.

Depending on the application, many other factors can play a role. For instance, if the compression algorithm has to be implemented in an instrument on a satellite where it can only draw on very limited power supplies, then it is also important for the computations involved in the transform itself to be as economical as possible.

Readers who want to know more about (important!) considerations of this kind can find them discussed in the engineering literature. Readers who are content to stay at the lofty mathematical level are of course welcome to do so, but are hereby warned that there is more to image compression via wavelet transforms than has been sketched in the previous sections.

7 Brief Overview of Several Influences on the Development of Wavelets

Most of what is now called “wavelet theory” was developed in the 1980s and early 1990s. It built on existing work and insights from many fields, including harmonic analysis (mathematics), computer vision and computer graphics (computer science), signal analysis and signal compression (electrical engineering), coherent states (theoretical physics), and seismology (geophysics). These different strands did not come together all at once but were brought together gradually, often as the result of serendipitous circumstances and involving many different agents.

In harmonic analysis, the roots of wavelet theory go back to work by LITTLEWOOD [VI.79] and Paley in the 1930s. An important general principle in Fourier analysis is that the smoothness of a function is reflected in its FOURIER TRANSFORM [III.27]: the smoother the function, the faster the decay of its transform. Littlewood and Paley addressed the question of characterizing *local* smoothness. Consider, for example, a periodic function with period 1 that has just one discontinuity in the interval $[0, 1)$ (which is then repeated at all integer translates of that point), and is smooth elsewhere. Is the smoothness reflected in the Fourier transform?

If the question is understood in the obvious way, then the answer is no: a discontinuity causes the Fourier coefficients to decay slowly, however smooth the rest of the function is. Indeed, the best possible decay is of the form $|\hat{f}_n| \leq C[1 + |n|]^{-1}$. If there were no discontinuity, the decay would be at least as good as $C_k[1 + |n|]^{-k}$ when f is k -times differentiable.

However, there is a more subtle connection between local smoothness and Fourier coefficients. Let f be a periodic function, and let us write its n th Fourier coefficient \hat{f}_n as $a_n e^{i\theta_n}$, where a_n is the absolute value of \hat{f}_n and $e^{i\theta_n}$ is its *phase*. When examining the decay of the Fourier coefficients, we look just at a_n and forget all about the phases, which means that we cannot detect any phenomenon unless it is unaffected by arbitrary changes to the phases. If f has a discontinuity, then we can clearly move it about by changing the phases. It turns out that these phases play an important role in determining not just where the singularities are, but even their severity: if the singularity at x_0 is not just a discontinuity but a divergence of the type $|f(x)| \sim |x - x_0|^{-\beta}$, then one can change the value of β just by changing the phases and without altering the absolute values $|a_n|$. Thus, changing phases in Fourier series is a dangerous thing to do: it can greatly change the properties of the function in question.

Littlewood and Paley showed that *some* changes of the phases of Fourier coefficients are more innocuous. In particular, if you choose a phase change for the first Fourier coefficient, another one for both the next two coefficients, another for the next four, another for the next eight, and so on, so that the phase changes are constant on “blocks” of Fourier coefficients that keep doubling in length, then local smoothness (or absence of smoothness) properties of f are preserved. Similar statements hold for the Fourier transform of functions on \mathbb{R} (as opposed to Fourier series of periodic functions). This was the first result of a whole branch of harmonic analysis in which *scaling* was exploited systematically to deal with detailed local analysis, and in which very powerful theorems were proved that, with hindsight, seem ready-made to establish a host of powerful properties for wavelet decompositions. The simplest way to see the connection between Littlewood-Paley theory and wavelet decompositions is to consider the *Shannon wavelet* $\psi^{[\text{Sh}]}$, which is defined by $\hat{\psi}^{[\text{Sh}]}(\xi) = 1$ when $\pi \leq |\xi| < 2\pi$, and $\hat{\psi}^{[\text{Sh}]}(\xi) = 0$ otherwise. Here, $\hat{\psi}^{[\text{Sh}]}$ denotes the Fourier transform of the wavelet $\psi^{[\text{Sh}]}$. The corresponding functions $\psi_{j,k}^{[\text{Sh}]}(x) = 2^{j/2} \psi^{[\text{Sh}]}(2^j x - k)$ constitute an orthonormal basis for

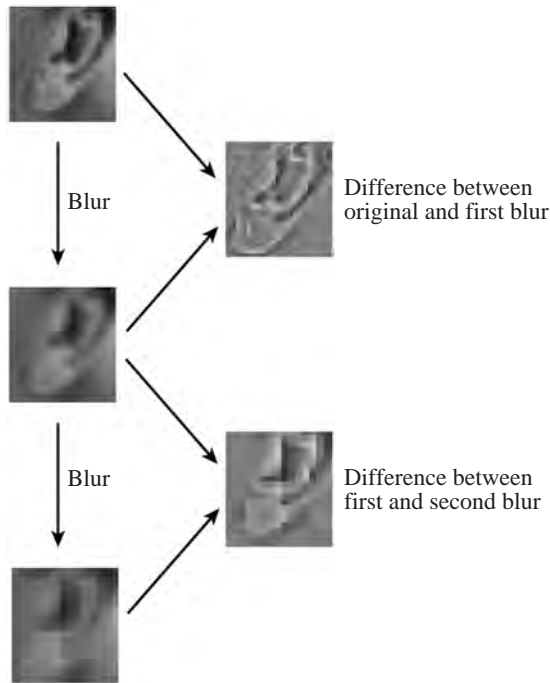


Figure 12 Differences between successive blurs give detail at different scales.

$L^2(\mathbb{R})$, and for each f and each j the collection of inner products $(\int_{-\infty}^{\infty} f(x) \psi_{j,k}^{[Sh]}(x) dx)_{k \in \mathbb{Z}}$ tells us how $\hat{f}(\xi)$ restricts to the set $2^{j-1} \leq \pi^{-1}|\xi| < 2^j$. In other words, it gives us the j th Littlewood-Paley block of f .

Scaling also plays an important role in computer vision, where one of the basic ways to “understand” an image (going back to at least the early 1970s) is to blur it more and more, erasing more detail each time, so as to obtain approximations that are graded in “coarseness” (see figure 12). Details at different scales can then be found by considering the differences between successive coarsenings. The relationship with wavelet transforms is obvious!

An important class of signals of interest to electrical engineers is that of *bandlimited signals*, which are functions f , usually of one variable only, for which the Fourier transform \hat{f} vanishes outside some interval. In other words, the frequencies that make up f come from some “limited band.” If the interval is $[-\Omega, \Omega]$, then f is said to have *bandlimit* Ω . Such functions are completely characterized by their values, often called *samples*, at integer multiples of π/Ω . Most manipulations on the signal f are carried out not directly but by operations on this sequence of samples. For instance, we

might want to restrict f to its “lower-frequency half.” To do this, we would define a function g by the condition that $\hat{g}(\xi) = \hat{f}(\xi)$ if $|\xi| \leq \Omega/2$ and is 0 otherwise. Equivalently, we could say that $\hat{g}(\xi) = \hat{f}(\xi)\hat{L}(\xi)$, where $\hat{L}(\xi) = 1$ if $|\xi| \leq \Omega/2$ and 0 otherwise. The next step is to let L_n be $L(n\pi/\Omega)$, and we find that $g(k\pi/\Omega) = \sum_{n \in \mathbb{Z}} L_n f((k-n)\pi/\Omega)$. To put this more neatly, if we write a_n and \tilde{b}_n for $f(n\pi/\Omega)$ and $g(n\pi/\Omega)$, respectively, then $\tilde{b}_k = \sum_{n \in \mathbb{Z}} L_n a_{k-n}$. On the other hand, g clearly has bandlimit $\Omega/2$, so to characterize g it suffices to know only the sequence of samples at integer multiples of $2\pi/\Omega$. In other words, we just need to know the numbers $b_k = \tilde{b}_{2k}$. The transition from f to g is therefore given by $b_k = \sum_{n \in \mathbb{Z}} L_n a_{2k-n}$. In the appropriate electrical engineering vocabulary, we have gone from a critically sampled sequence for f (i.e., its sampling rate corresponded exactly to its bandlimit) to a critically sampled sequence for g by *filtering* (multiplying \hat{f} by some function, or convolving the sequence $(f(n\pi/\Omega))_{n \in \mathbb{Z}}$ with a sequence of *filter coefficients*) and *downsampling* (retaining only one sample in two, because these are the only samples necessary to characterize the more narrowly bandlimited g). The upper-frequency half h of f can be obtained by the inverse Fourier transform of the restriction of $\hat{f}(\xi)$ to $|\xi| > \Omega/2$. Like g , the function h is also completely characterized by its values at multiples of $2\pi/\Omega$, and h can also be obtained from f by filtering and downsampling. This split of f into its lower and upper frequency halves, or *subbands*, is thus given by formulas that are the exact equivalent of the generalized averaging and differencing encountered in the implementation of wavelet transforms for orthonormal wavelet bases supported on an interval. Subband filtering followed by critical downsampling had been developed in the electrical engineering literature before wavelets came along, but were typically not concatenated in several stages.

A concept of central importance in quantum physics is that of a UNITARY REPRESENTATION [IV.15 §1.4] of a LIE GROUP [III.50 §1] on some HILBERT SPACE [III.37]. In other words, given a Lie group G and a Hilbert space H , one interprets the elements g of G as unitary transformations of H . The elements of H are called *states*, and for certain Lie groups, if \mathbf{v} is some fixed state, then the family of vectors $\{g\mathbf{v}; g \in G\}$ is called a *family of coherent states*. Coherent states go back to work by Schrödinger in the 1920s. Their name dates back to the 1950s, when they were used in quantum optics: the word “coherent” referred to the coherence of the

Terri: Tim thinks this sentence is fine. OK?

light they were describing. These families turned out to be of interest in a much wider range of settings in quantum physics, and the name stuck, even outside the original setting of optics. In many applications it helps not to use the whole family of coherent states but only those coherent states that correspond to a certain kind of discrete subset of \mathcal{G} . Wavelets turn out to be just such a subfamily of coherent states: one starts with a single, basic wavelet, and the transformations that convert it (by dilation and translation) into the remaining wavelets form a discrete semigroup of such transformations.

Despite the fact that wavelets synthesized ideas from all these fields, their discovery originated in another area altogether. In the late 1970s, the geophysicist J. Morlet was working for an oil company. Dissatisfied with the existing techniques for extracting special types of signals from seismograms, he came up with an ad hoc transform that combined translations and scalings: nowadays, it would be called a redundant wavelet transform. Other transforms in seismology with which Morlet was familiar involve comparing the seismic traces with special functions of the form $W_{m,n}(t) = w(t - n\tau) \cos(m\omega t)$, where w is a smooth function that gently rises from 0 to 1 and then gently decays to 0 again, all within a finite interval. Several different examples of functions w , proposed by several different scientists, are used in practice: because the functions $W_{m,n}$ look like small waves (they oscillate, but have a nice beginning and end because of w) they are typically called “wavelets of X,” named after proposer X for that particular w . The reference functions in Morlet’s new ad hoc family, which he used to compare pieces of seismic traces, were different in that they were produced from a function w by *scaling* instead of multiplying them by increasingly oscillating trigonometric functions. Because of this, they always had the same shape, and Morlet called them “wavelets of constant shape” (see figure 13) in order to distinguish them from the wavelets of X (or Y, or Z, etc.).

Morlet taught himself to work with this new transform and found it numerically useful, but had difficulty explaining his intuition to others because he had no underlying theory. A former classmate pointed him in the direction of A. Grossmann, a theoretical physicist, who made the connection with coherent states and, together with Morlet and other collaborators, started to develop a theory for the transform in the early 1980s. Outside the field of geophysics it was no longer necessary to use the phrase “of constant shape,” so this was

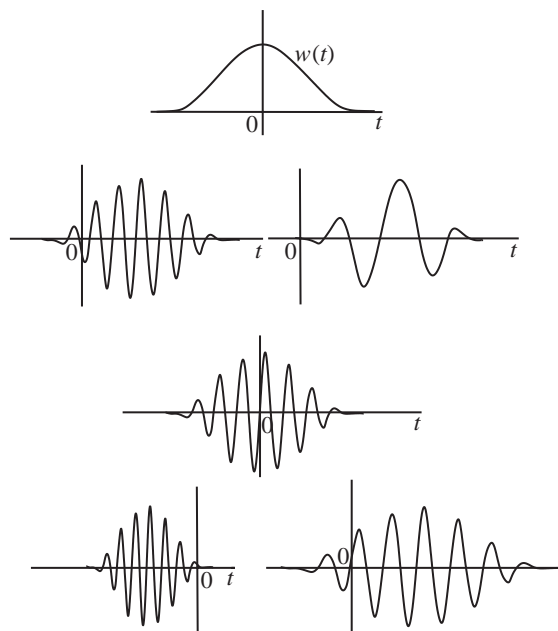


Figure 13 Top: an example of a window function w that is used in practice by geophysicists, with just below it two examples of $w(t - n\tau)e^{imt}$, i.e., two “traditional” geophysics wavelets. Bottom: a wavelet as used by Morlet, with two translates and dilates just below it—these have constant shape, unlike the “traditional” ones.

quickly dropped, which annoyed geophysicists when, some years later, more mature forms of wavelet theory impinged on their field again.

A few years later, in 1985, standing in line for a photocopy machine at his university, harmonic analysis expert Y. Meyer heard about this work and realized it presented an interestingly different take on the scaling techniques with which he and other harmonic analysts had long been familiar. At the time, no wavelet bases were known in which the initial function ψ combined the properties of smoothness and good decay. Indeed, there seemed to be a subliminal expectation in papers on wavelet expansions that no such orthonormal wavelet bases could exist. Meyer set out to prove this, and to everyone’s surprise and delight he failed in the best possible way—by finding a counterexample, the first smooth wavelet basis! Except that it later turned out not to have been the very first: a few years before, a different harmonic analyst, O. Stromberg, had constructed a different example, but this had not attracted attention at the time.

Terri: this term is indeed OK.

Meyer's proof was ingenious, and worked because of some seemingly miraculous cancellations, which is always unsatisfactory from the point of view of mathematical understanding. Similar miracles played a role in independent constructions by P. G. Lemarié (now Lemarié-Rieusset) and G. Battle of orthonormal wavelet bases that were piecewise polynomial. (They came to the same result from completely different points of departure—harmonic analysis for Lemarié and quantum field theory for Battle.)

A few months later, S. Mallat, then a Ph.D. candidate in computer vision in the United States, learned about these wavelet bases. He was on vacation, chatting on the beach with a former classmate who was one of Meyer's graduate students. After returning to his Ph.D. work, Mallat kept thinking about a possible connection with the reigning paradigm in computer vision. On learning that Meyer was coming to the United States in the fall of 1986 to give a named lecture series, he went to see him and explain his insight. In a few days of feverish enthusiasm, they hammered out *multiresolution analysis*, a different approach to Meyer's construction inspired by the computer vision framework. In this new setting, all the miracles fell into place as inevitable consequences of simple, entirely natural construction rules, embodying the principle of successively finer approximations. Multiresolution analysis has remained the basic principle behind the construction of many wavelet bases and redundant families.

None of the smooth wavelet bases constructed up to that point was supported inside an interval, so the algorithms to implement the transform (which were using the subband filtering framework without their creators knowing that it had been named and developed in electrical engineering) required, in principle, infinite filters that were impossible to implement. In practice, this meant that the infinite filters from the mathematical theory had to be truncated; it was not clear how to construct a multiresolution analysis that would lead to finite filters. Truncation of the infinite filters seemed to me a blemish on the whole beautiful edifice, and I was unhappy with this state of affairs. I had learned about wavelets from Grossmann and about multiresolution analysis from explanations scribbled by Meyer on a napkin after dinner during a conference. In early 1987 I decided to insist on finite filters for the implementation. I wondered whether a whole multiresolution analysis (and its corresponding orthonormal basis of wavelets) could be reconstructed from appropriate but finite filters. I managed to carry out this pro-

gram, and as a result found the first construction of an orthonormal wavelet basis for which ψ is smooth and supported on an interval.

Soon after this, the connection with the electrical engineering approaches was discovered. Especially easy algorithms were inspired by the needs of computer graphics applications. More exciting constructions and generalizations followed: biorthogonal wavelet bases, wavelet packets, multiwavelets, irregularly spaced wavelets, sophisticated multidimensional wavelet bases not derived from one-dimensional constructions, and so on.

It was a heady, exciting period. The development of the theory benefited from all the different influences and also enriched the different fields with which wavelets are related. As the theory has matured, wavelets have become an accepted addition to the mathematical toolbox used by mathematicians, scientists, and engineers alike. They have also inspired the development of other tools that are better adapted to tasks for which wavelets are not optimal.

Further Reading

- Aboufadel, E., and S. Schlicker. 1999. *Discovering Wavelets*. New York: Wiley Interscience.
- Blatter, C. 1999. *Wavelets: A Primer*. Wellesley, MA: AK Peters.
- Cipra, B. A. 1993. Wavelet applications come to the fore. *SIAM News* 26(7):10–11, 15.
- Frazier, M. W. 1999. *An Introduction to Wavelets through Linear Algebra*. New York: Springer.
- Hubbard, B. B. 1995. *The World According to Wavelets: The Story of a Mathematical Technique in the Making*. Wellesley, MA: AK Peters.
- Meyer, Y., and R. Ryan. 1993. *Wavelets: Algorithms and Applications*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).
- Mulcahy, C. 1996. Plotting & scheming with wavelets. *Mathematics Magazine* 69(5):323–43.

VII.4 The Mathematics of Traffic in Networks

Frank Kelly

1 Introduction

We are all familiar with congested roads, and perhaps also with congestion in other networks such as the Internet, so it is obviously important to have a general understanding of how and why congestion occurs

in networks. However, the pattern of the flow of traffic through a network is the consequence of a subtle and complex interaction between different users. For example, in a road network we would normally expect each driver to attempt to choose the most convenient route, and this choice will depend upon the delays the driver expects to encounter on different roads; but these delays will in turn depend upon the choices of routes made by others. This mutual interdependence makes it difficult to predict the effects of changes to the system, such as the construction of a new road or the introduction of tolls in certain places.

Related issues arise in other large-scale systems like the telephone network or the Internet. In these systems a major practical concern is the extent to which control can be *decentralized*. When you are browsing the Web, the rate at which a Web page is transferred to you across the network is controlled by software protocols running on your computer and on the Web server hosting the Web page, and not by some huge central computer. This decentralized approach to flow control has been outstandingly successful as the Internet has evolved from a small-scale research network to today's interconnection of hundreds of millions of hosts, but is beginning to show signs of strain. In developing new protocols, the challenge is to understand just which aspects of decentralized flow control are important if the network as a whole is to continue to expand and evolve.

In this article we introduce the reader to some of the mathematical models that have been used to address these issues. The models need to be able to represent several distinct aspects of the system. We shall see that the language of GRAPH THEORY [III.34] and MATRICES [I.3 §4.2] is needed to capture the pattern of connections within the network. Calculus is needed to describe how congestion depends upon traffic volumes. And optimization concepts are needed to model the way in which self-interested drivers choose their shortest routes, or the way that decentralized controls in communication networks can cause the system as a whole to perform well.

2 Network Structure

Figure 1 illustrates a set of three nodes connected by a set of five directed links. We might imagine the nodes as representing towns or locations within a city, and the links as representing road capacity between different nodes. A two-way road is represented by two links, one

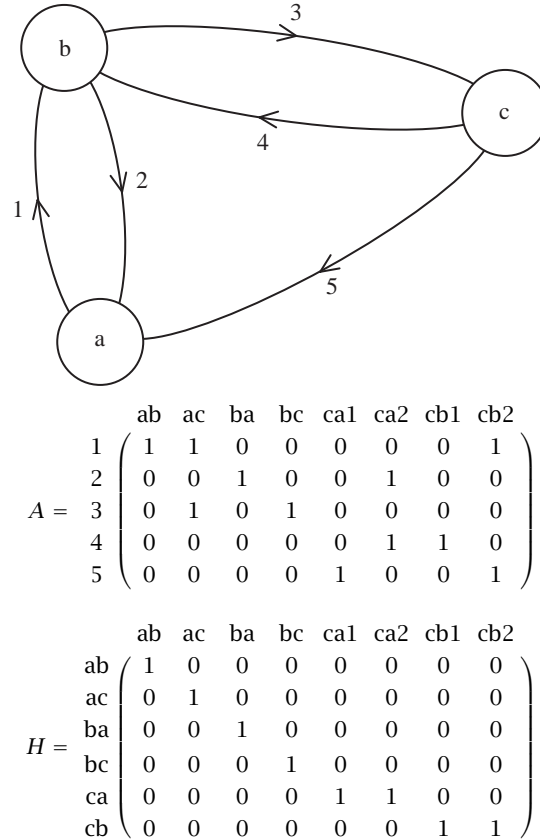


Figure 1 A simple network and its link-route incidence matrix, A . The matrix H represents which routes serve which source-destination pairs.

in each direction. Notice that there are two routes from node c to node a that a driver can choose: the first route, let us call it ca1, is the direct route, using link 5; the second route, let us call it ca2, is via node b and uses links 4 and 2.

Let J be the set of directed links and let R be the set of possible routes. One way to describe the relationship between links and routes is with a table, or *matrix*, defined as follows. Set $A_{jr} = 1$ if link j lies on route r , and set $A_{jr} = 0$ otherwise. This defines a matrix $A = (A_{jr}, j \in J, r \in R)$ called the *link-route incidence matrix*. Each column of the matrix corresponds to one of the routes r , and each row to one of the links j of the network. The column for route r is composed of 0s and 1s: the 1s tell us which links are on route r . As for the rows, the 1s in the row for link j tell us which routes pass through that link. Thus, for example, the incidence matrix in figure 1 has a column

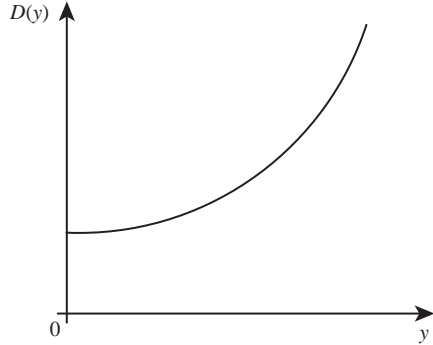


Figure 2 The time taken to travel along a link, $D(y)$, expressed as a function of the total flow y along the link. As the flow increases, congestion effects cause additional delay.

for each of the two routes, ca1 and ca2, between node c and node a. These columns encode the information that route ca1 uses link 5 and that route ca2 uses links 4 and 2. Note that the incidence matrix does not tell us the order of the links on the route. Also the incidence matrix shown does not include all logically possible routes, but it could if we wanted it to. And while we have illustrated a very small network, there is no limit to the number of nodes and links there could be in the network, or to the number of choices of route each driver might have—the incidence matrix would just be bigger.

One quantity of interest in a network is the volume of traffic along a particular route or link. Let x_r be the *flow* on route r , defined as the number of cars per hour that travel along that route. We can list the flows along all the routes in the network as a sequence of numbers $x = (x_r, r \in R)$, and we can think of this sequence as a vector. From this vector we can calculate the total flow through a link: for example, the total flow through link 5 in figure 1 is the sum of the flows along routes ca1 and cb2, since these are the routes that pass through link 5. In general, since $A_{jr} = 1$ when a route r passes through link j and $A_{jr} = 0$ when it does not, the total flow through link j , coming from all of the routes that use it, is

$$y_j = \sum_{r \in R} A_{jr} x_r, \quad j \in J.$$

Again, the numbers $(y_j, j \in J)$ can be thought of as forming a vector. The above equations can then be represented succinctly in matrix form as

$$y = Ax.$$

We expect the level of congestion at a link to depend on the total flow through the link, and we expect this to influence the time taken to travel along the link. We shall call this time the *delay*. Figure 2 shows a typical way in which the delay might depend on the amount of flow. At small values of the flow y the delay $D(y)$ is just the time taken to travel along an empty road; for larger values of y the delay $D(y)$ is larger, and quite possibly *much* larger, owing to congestion effects.¹

Let $D_j(y_j)$ be the delay along link j when the flow through that link is y_j ; the nature of this delay may depend upon characteristics of link j such as its length and width, so we have to use the subscript j on the function D_j to indicate that the functions for the various links can be different.

2.1 Routing Choices

Given two nodes in a network there will in general be a variety of possible routes capable of linking them. For example, in figure 1 we have seen that the incidence matrix A records two routes between nodes c and a. The pair ca is an example of a *source-destination pair*. Flow originating from source c and destined for node a can use either ca1 or ca2, the two routes that serve this source-destination pair. We now need another matrix, this time to describe the relationship between source-destination pairs and routes. Let us use s to denote a typical source-destination pair, and let S be the set of all source-destination pairs. Then, for each source-destination pair s and each route r , let $H_{sr} = 1$ if s can be served by the route r , and let $H_{sr} = 0$ otherwise. This defines a matrix $H = (H_{sr}, s \in S, r \in R)$; figure 1 gives an example. Observe that the row labeled ca has 1s for the two routes, $r = \text{ca1}, \text{ca2}$, that serve the source-destination pair $s = \text{ca}$. Each column of H corresponds to a route, and contains a single 1: this identifies the source-destination pair served by the route. For each route r let us write $s(r)$ for the source-destination pair served by r : for example, in figure 1, $s(\text{ac}) = \text{ac}$ and $s(\text{ca1}) = \text{ca}$.

1. The graph shown in figure 2 is single valued. It is quite possible for the curve representing delay as a function of flow to bend back upon itself, so that higher delays than shown in the graph correspond to flows *smaller* than the maximum flow shown there. You are in this part of the graph when you experience stop-start driving conditions on a congested but otherwise incident-free highway. Part of the aim of traffic management is to keep flows and delays away from this part of the graph, which we will not consider further.

We will assume that the graph is increasing and smooth, which will make our use of calculus later more straightforward. Formally, we shall assume that $D(y)$ is a continuously differentiable and strictly increasing function of its argument y , as in the graph shown in figure 2.

From the vector $x = (x_r, r \in R)$ we can calculate the total flow from a source to a destination: for example, the flow from node c to node a in figure 1 is the sum of flows along routes $ca1$ and $ca2$, since from the matrix H we see that these are the routes that serve the source-destination pair ca . More generally, if f_s is the total flow of traffic added up over all of the routes serving source-destination pair s , then

$$f_s = \sum_{r \in R} H_{sr} x_r, \quad s \in S.$$

Thus the vector $f = (f_s, s \in S)$ of source-destination flows can be expressed succinctly in matrix form as $f = Hx$.

3 Wardrop Equilibria

We are now able to approach the central issue: how do the traffic flows between the various sources and destinations distribute themselves over the links of the network? Each driver will try to use whatever route is quickest, but this may make other routes quicker or slower and cause other drivers to change their routes. Only when they cannot find alternative, quicker routes will drivers not have an incentive to change routes. What does this mean mathematically?

Let us first calculate the time taken for a driver to travel along route r . The column labeled r of the matrix A tells us which links j are on route r . If we add up the delays on each of these links, we get the time taken to travel along route r as the expression

$$\sum_{j \in J} D_j(y_j) A_{jr}.$$

Now the driver using route r could have used any other route that served the same source-destination pair $s(r)$. So, for the driver to be content with route r , we require

$$\sum_{j \in J} D_j(y_j) A_{jr} \leq \sum_{j \in J} D_j(y_j) A_{jr'},$$

for every other route r' that serves the same source-destination pair $s(r)$.

Define a *Wardrop equilibrium* (Wardrop 1952) to be a vector $x = (x_r, r \in R)$ of nonnegative numbers such that for every pair of routes r, r' serving the same source-destination pair,

$$x_r > 0 \Rightarrow \sum_{j \in J} D_j(y_j) A_{jr} \leq \sum_{j \in J} D_j(y_j) A_{jr'},$$

where $y = Ax$. The inequality expresses the defining characteristic of a Wardrop equilibrium: that if a route

r is actively used, then it achieves the minimum delay over all routes serving its source-destination pair $s(r)$.

Does a Wardrop equilibrium exist? It is not at all clear whether it is possible to find a vector x such that all of the above inequalities, for the various routes through the network, are satisfied simultaneously. To answer the question, we shall proceed by addressing a seemingly different question: what is the answer to the following optimization problem?

$$\begin{aligned} &\text{Minimize} \quad \sum_{j \in J} \int_0^{y_j} D_j(u) du \\ &\text{over} \quad x \geq 0, \quad y, \\ &\text{subject to} \quad Hx = f, \quad Ax = y. \end{aligned}$$

Let us see in outline why this optimization problem has a solution (x, y) , and why, if (x, y) is a solution, the vector x is a Wardrop equilibrium.

The optimization problem has some aspects that are quite natural. An obvious constraint is that the flows along each route are nonnegative, which is why we insist that $x \geq 0$. The constraints $Hx = f, Ax = y$ just enforce the accounting rules we have seen earlier—the rules that allow the source-destination flows f and the link flows y to be calculated from the route flows x using the matrices H and A , respectively. We view the source-destination flows f as fixed, to be distributed over the various routes. Given a choice of f , our task is then to find the route flows x and consequently the link flows y . At a solution to the optimization problem y will be nonnegative, since x is.

This much is fairly natural, but the function to be minimized looks somewhat strange. Its importance rests on the fact that the rate of change of the integral

$$\int_0^{y_j} D_j(u) du$$

with respect to y_j is $D_j(y_j)$, by THE FUNDAMENTAL THEOREM OF CALCULUS [I.3 §5.5], and the function to be minimized is the sum of these integrals over all links. We shall see that the link between Wardrop equilibria and the optimization problem is a direct consequence of this observation.

To find a solution to the optimization problem, we will use the method of LAGRANGE MULTIPLIERS [III.66]. Define the function

$$\begin{aligned} L(x, y; \lambda, \mu) \\ = \sum_{j \in J} \int_0^{y_j} D_j(u) du + \lambda \cdot (f - Hx) - \mu \cdot (y - Ax), \end{aligned}$$

where $\lambda = (\lambda_s, s \in S)$, $\mu = (\mu_j, j \in J)$ are vectors of Lagrange multipliers, to be fixed later. The idea is that if we make the right choices of Lagrange multipliers, the minimization of the function L over x and y will find a solution to the original problem. The reason this works is that, for the right choices of Lagrange multipliers, the constraints $Hx = f$ and $Ax = y$ are consistent with the minimization of L .

To minimize the function L we need to differentiate. First,

$$\frac{\partial L}{\partial y_j} = D_j(y_j) - \mu_j.$$

Second,

$$\frac{\partial L}{\partial x_r} = -\lambda_{s(r)} + \sum_{j \in J} \mu_j A_{jr}.$$

Note that the form of the matrix H causes the derivative with respect to x_r to pick out exactly one component of λ , namely $\lambda_{s(r)}$, and the form of the matrix A causes the derivative to pick out just those components of μ that correspond to links on route r . These derivatives allow us to deduce that a minimum of L , over all $x \geq 0$ and all y , occurs when

$$\mu_j = D_j(y_j)$$

and

$$\begin{aligned} \lambda_{s(r)} &= \sum_{j \in J} \mu_j A_{jr} \quad \text{if } x_r > 0 \\ &\leq \sum_{j \in J} \mu_j A_{jr} \quad \text{if } x_r = 0. \end{aligned}$$

The equality condition for $\lambda_{s(r)}$ is straightforward: if $x_r > 0$ then small variations up or down in x_r should not decrease the function $L(x, y; \lambda, \mu)$, and hence we deduce that the partial derivative with respect to x_r must be zero. But if $x_r = 0$ then we can only vary x_r upward, and so all we can deduce is that the partial derivative with respect to x_r is nonnegative, and from this we deduce the inequality condition for $\lambda_{s(r)}$.

Minimizing the function L corresponds to allowing the constraints $Hx = f$, $Ax = y$ to be violated, but at a cost: now one charges a price λ_s for any shortfall of the sum $\sum_{j \in J} A_{jr} x_r$ below f_s and a price μ_j for any excess of the sum $\sum_{j \in J} A_{jr} x_r$ over y_j . From general results on convex optimization it is known that there exist Lagrange multipliers (λ, μ) and a vector (x, y) such that (x, y) minimizes $L(x, y; \lambda, \mu)$, satisfies the constraints $Hx = f$, $Ax = y$, and solves the original optimization problem.

Our solution for the Lagrange multipliers shows that they have a simple interpretation: μ_j is the delay on link

j and λ_s is the minimum delay over all routes serving the node pair s . The various conditions established for the multipliers thus show that an optimum of the function L , known as the *objective function*, corresponds precisely to a Wardrop equilibrium.

Thus if traffic in the network distributes itself in accordance with the self-interested choices of drivers, the equilibrium flows (x, y) will solve an optimization problem. This result is originally due to Beckmann et al. (1956), and it provides a remarkable insight into the equilibrium patterns achieved in road traffic networks. The pattern of traffic resulting from the individual decisions of a large number of self-interested drivers behaves as if a central intelligence were directing flows to optimize a certain (rather strange) objective function.

The result does *not* mean that average delays in the network will be minimal: a striking illustration of this fact is provided by Braess's paradox (Braess 1968), which we describe next.

4 Braess's Paradox

Consider the network illustrated in figure 3(a). Cars travel from node S to node N, via either node W or node E. The total flow is 6, and the link delays $D_j(y)$ are given next to the links in the figure. One can imagine the figure illustrating rush hour as commuters travel from the center of a city in the south to their homes in the north. Commuters learn from experience what the delays are likely to be along the eastern and western routes. The distribution of traffic shown is the Wardrop equilibrium: there is no incentive for any drivers to change their routes, since the two possible routes incur the same delay, namely $(10 \times 3) + (3 + 50) = 83$ units of time. Now suppose that a new link is added, between nodes W and E, as shown in figure 3(b). Traffic is attracted onto the new link, since to begin with it offers a shorter journey time from the south to the north. Eventually, after everyone knows about the new link and traffic patterns have settled down, a new Wardrop equilibrium will be established, and this is shown in figure 3(b). In the new equilibrium there are three routes used, which each incur the same delay, namely $(10 \times 4) + (2 + 50) = (10 \times 4) + (2 + 10) + (10 \times 4) = 92$. Thus in figure 3(b) each car incurs a delay of 92, while in figure 3(a) the delay of each car was only 83. Adding the new link has increased everyone's delay!

The explanation for this apparent paradox is as follows. At a Wardrop equilibrium each driver is using

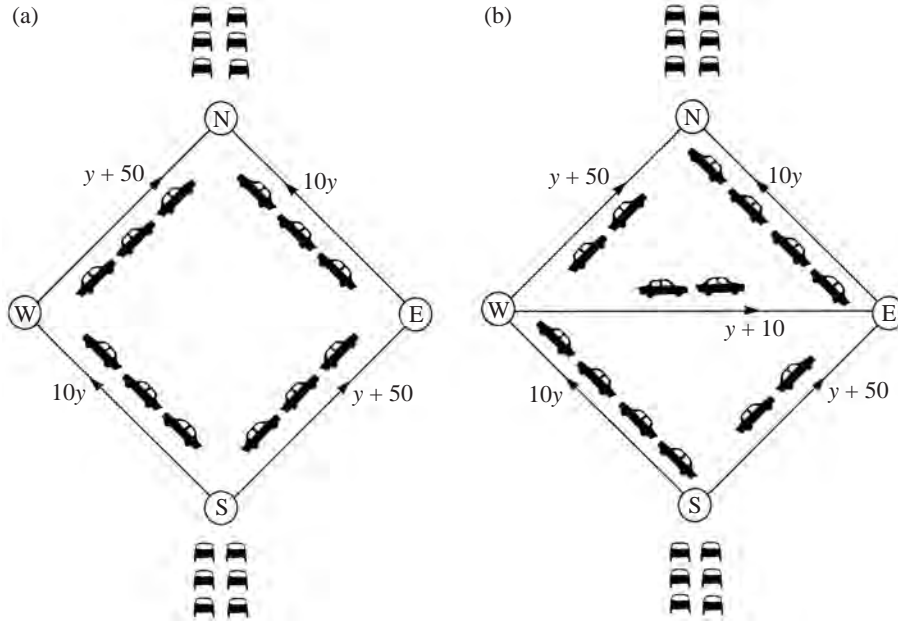


Figure 3 Braess's paradox. The addition of a link causes everyone's journey time to lengthen. (After Braess (1968) and Cohen (1988).)

a route which, given the choices of others, gives the minimum delay over the routes available between that driver's source and destination. But there is no intrinsic reason why this equilibrium should correspond to particularly low delays relative to what could be achieved by another flow pattern. If all drivers could be encouraged to depart from their own self-interested choices, it is quite possible that all might benefit. And in the above example, if all drivers in the second network could agree to avoid the new link, effectively converting the network back into the first network, then all would incur lower delays.

To explore the point further, note that the product of the flow y_j and the delay $D_j(y_j)$ is the delay incurred at link j per unit time, aggregated over all the vehicles using link j . Let us try to find the flow pattern that minimizes the total delay per unit time, summed over the entire network. Consider then the following problem.

$$\begin{aligned} & \text{Minimize} && \sum_{j \in J} y_j D_j(y_j) \\ & \text{over} && x \geq 0, y, \\ & \text{subject to} && Hx = f, Ax = y. \end{aligned}$$

Note that the problem is of the same form as the earlier optimization problem, but the function to be minimized now measures the total network delay per unit

time. (Recall that the function to be minimized in the first optimization problem seemed initially to be rather arbitrary, with its eventual motivation being that its minimization was achieved by a Wardrop equilibrium.) Again define the function

$$\begin{aligned} L(x, y; \lambda, \mu) &= \sum_{j \in J} y_j D_j(y_j) + \lambda \cdot (f - Hx) - \mu \cdot (y - Ax). \end{aligned}$$

Again

$$\frac{\partial L}{\partial x_r} = -\lambda_{s(r)} + \sum_{j \in J} \mu_j A_{jr},$$

but now

$$\frac{\partial L}{\partial y_j} = D_j(y_j) + y_j D'_j(y_j) - \mu_j.$$

Hence a minimum of L over $x \geq 0$ and y occurs when

$$\mu_j = D_j(y_j) + y_j D'_j(y_j)$$

and

$$\begin{aligned} \lambda_{s(r)} &= \sum_{j \in J} \mu_j A_{jr} && \text{if } x_r > 0 \\ &\leq \sum_{j \in J} \mu_j A_{jr} && \text{if } x_r = 0. \end{aligned}$$

The Lagrange multipliers now have a more sophisticated interpretation. Suppose that, in addition to the

delay $D_j(y_j)$, users of link j incur a traffic-dependent toll

$$T_j(y_j) = y_j D'_j(y_j).$$

Then μ_j is the *generalized cost* of using link j , defined as the sum of the toll and the delay, and λ_s is the minimum generalized cost over all routes serving the node pair s . If users select routes in an attempt to minimize the sum of their tolls and their delays, then they will produce a flow pattern that minimizes total delay in the network. Notice that the generalized cost μ_j is $(\partial/\partial y_j)(y_j D(y_j))$, which is the rate of increase in the total delay at link j as the flow y_j is increased. So the assumption now is that, in a certain sense, drivers try to minimize their contribution to the total delay rather than minimizing their own delay.

We have seen that if drivers attempt to minimize their own delay, then the resulting equilibrium flows will minimize a certain objective function defined for the network. However, the objective function is certainly not the total network delay, and thus there is no guarantee that when capacity is added to a network the situation is improved. We have also seen that, with the imposition of appropriate tolls, it is possible for the self-interested behavior of drivers to lead to an equilibrium pattern of flow that minimizes total delay. A major challenge for governments and transport planners is to understand how insights from these and more sophisticated models might be used to encourage more efficient development and use of road networks (Department for Transport 2004).

5 Flow Control in the Internet

When a file is requested over the Internet, the computer that hosts that file breaks it into small packets of data that are then transferred across the network by the *transmission control protocol* of the Internet, or TCP. The rate at which packets enter the network is controlled by TCP, which is implemented as software on the two computers that are the source and destination of the data. The general approach is as follows (Jacobson 1988). When a link within the network becomes overloaded, one or more packets are lost; loss of a packet is taken as an indication of congestion, the destination informs the source, and the source slows down. TCP then gradually increases its sending rate until it again receives an indication of congestion. This cycle of increase and decrease enables the source computers to discover and use the available capacity, and to share it between different flows of packets.

TCP has been outstandingly successful as the Internet has evolved from a small-scale research network to today's interconnection of hundreds of millions of endpoints and links. This in itself is a striking observation. Each of a large but indeterminate number of flows is controlled by a feedback loop that can know only of that flow's experience of congestion. A flow does not know how many other flows are sharing a link on its route, or even how many links are on its route. The links vary in capacity by many orders of magnitude, as do the numbers of flows sharing different links. It is remarkable that so much has been achieved in such a rapidly growing and heterogeneous network with congestion controlled just at the endpoints. Why does this algorithm work so well?

In recent years theoreticians have shed some light on TCP's success, by interpreting the protocol as a decentralized parallel algorithm that solves an optimization problem, just as the decentralized choices of drivers in a road network solve an optimization problem. We shall outline the argument, beginning with a more detailed description of TCP.²

Packets transferred by TCP across the Internet contain *sequence numbers* indicating their order, and they should arrive at their destination in that order. When a packet is received at the destination, it is acknowledged: an acknowledgment is a short packet sent by the destination back to the source. If a packet has been lost in the transfer, the source can tell this from the sequence numbers contained in the acknowledgments. The source keeps a copy of each packet sent until it has been positively acknowledged; these copies form what is called a *sliding window*, and allow packets lost in transfer to be sent again by the source.

Meanwhile, stored in the source computer there is a numerical variable known as the *congestion window* and denoted *cwnd*. The congestion window directs the size of the sliding window in the following sense: if the size of the sliding window is less than *cwnd*, then the computer increases it by sending out a packet; if it is greater than or equal to *cwnd*, then it waits for positive acknowledgments to come in, which have the effect of reducing the size of the sliding window and, as we shall see, increasing *cwnd* as well. Thus, the size of the sliding window continually changes, moving in the direction of a target size that is given by the congestion window.

2. Even our detailed description of TCP is simplified, concerning just the congestion-avoidance part of the protocol and omitting discussion of timeouts or of reactions to multiple congestion indication signals received within a single round-trip time.

The congestion window itself is not a fixed number: rather, it is constantly being updated, and the precise rules for how this is done are critical for TCP's sharing of capacity. The rules currently used are as follows. Every time a positive acknowledgment comes in, cwnd is increased by cwnd^{-1} , and every time a lost packet is detected, cwnd is halved.³ Thus, if the source computer detects a lost packet, it realizes that there has been some congestion and backs off for a while, but if all its packets are getting through then it allows the rate at which it sends packets to inch up again.

If p is the probability that a packet is lost, then with probability $1 - p$ the congestion window will increase by cwnd^{-1} and with probability p it will decrease by $\frac{1}{2}\text{cwnd}$. The expected change in the congestion window cwnd per update step is therefore

$$\text{cwnd}^{-1}(1 - p) - \frac{1}{2}\text{cwnd} p.$$

The expected change will be positive for small values of cwnd , but will become negative if cwnd is big enough. We might therefore expect an equilibrium for cwnd to arise when the expression is zero: that is, when

$$\text{cwnd} = \sqrt{\frac{2(1 - p)}{p}}.$$

Now let us see how this calculation can be extended to networks. Suppose that a network consists of a set of nodes connected by directed links, like the network illustrated in figure 1. As earlier, let J be the set of directed links, let R be the set of routes, and let $A = (A_{jr}, j \in J, r \in R)$ be the link-route incidence matrix. When a request reaches a computer in this network, that computer will set up a congestion window for the flow of packets that will result. Since there will be many different such congestion windows, they need to be labeled, and it is convenient to label them with the route that will be used for the flow. (Exactly how these flows are routed is a complicated and important question, but one that we shall not discuss here.) So, for each route r that is being used, let cwnd_r be the congestion window for that route. Let T_r be the *round-trip time* for the route r : that is, the time between the sending out of a packet and the receiving of an acknowledgment for it.⁴ Finally, define a variable x_r to be cwnd_r / T_r .

Now at any given time the sliding window consists of those packets that have been sent but not acknowledged. Therefore, if a packet has just been acknowledged and its round-trip has taken time T_r , the sliding window consists of all packets sent out in the last T_r time units. Since the source computer is aiming for the number of such packets to be about cwnd_r , we can interpret x_r to be the rate at which packets are transferred over route r . Thus, the numbers x_r form a flow vector that is closely analogous to the traffic flow vector discussed earlier.

As we did then, let us define a vector $y = Ax$, so that y_j is the total flow through link j , obtained by summing x_r over each route r that passes through link j . Let p_j be the proportion of packets that are lost, or "dropped," at link j . We expect p_j to be related to y_j , the total flow through link j , as follows. If y_j is less than the capacity C_j of link j , then p_j will be zero; there will be no dropped packets at link j if the link is not full. And if $p_j > 0$ then $y_j = C_j$; if packets are dropped then the link is full. If we assume that the proportions of packets dropped at links are small, then the probability that a packet is lost on route r is approximately

$$p_r = \sum_{j \in J} p_j A_{jr}.$$

(The exact formula would be $(1 - p_r) = \prod_{j \in J} (1 - p_j)^{A_{jr}}$, but when the p_j are small we can ignore their products.) Since $x_r = \text{cwnd}_r / T_r$, our earlier calculation of cwnd now gives us that

$$x_r = \frac{1}{T_r} \sqrt{\frac{2(1 - p_r)}{p_r}}.$$

Is it possible to choose the rates $x = (x_r, r \in R)$ and the drop probabilities $p = (p_j, j \in J)$ in a consistent fashion, so that the last two equations are satisfied and either p_j is zero or $y_j = C_j$ for each $j \in J$? The remarkable observation is that such a choice corresponds precisely to the solution of the following optimization problem (Kelly 2001; Low et al. 2002).

$$\begin{aligned} & \text{Maximize} && \sum_{r \in R} \frac{\sqrt{2}}{T_r} \arctan \left(\frac{x_r T_r}{\sqrt{2}} \right) \\ & \text{over} && x \geq 0, \\ & \text{subject to} && Ax \leq C. \end{aligned}$$

3. These increase and decrease rules may appear rather mysterious, and indeed it is only recently that many of their macroscopic consequences have begun to be understood. The rules have worked well for more than a decade, but they are now beginning to show signs of age, and much current research is aimed at understanding the full consequences of changing them.

4. The round-trip time comprises the time taken for a packet to travel along links, called the propagation delay, together with pro-

cessing times and queueing delays at nodes. Processing times and queueing delays tend to decrease with increasing computer speeds, but the finite speed of light places a fundamental lower bound on propagation delays. We shall treat the round-trip time for a route as a constant. Hence, we assume that congestion at a link makes itself felt by packet loss rather than additional packet delay.

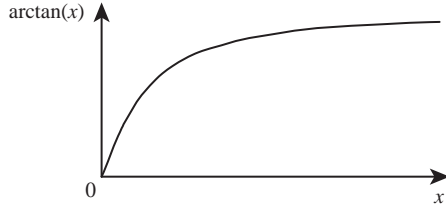


Figure 4 The arctan function. The Internet's TCP implicitly maximizes a sum of utilities over all the connections present in a network: this function shows the shape of the utility function for a single connection. The horizontal axis is proportional to the rate of the connection, and the vertical axis is proportional to the usefulness of that rate. Both axes are scaled in terms of the round-trip time of the connection.

Some aspects of this optimization problem are as we might expect: in particular, the inequality $Ax \leq C$ simply adds up the flows through link j and requires that the sum not exceed the capacity C_j of link j , for each link $j \in J$. But, as before, the function being optimized is undoubtedly strange. The arctan function, illustrated in figure 4, is the inverse function to the trigonometric function \tan , and can also be defined as

$$\arctan(x) = \int_0^x \frac{1}{1+u^2} du.$$

From this form, we see that its derivative with respect to x is $1/(1+x^2)$.

Let us sketch the relationship between the optimization problem and the equilibrium rates and drop probabilities. Define the function

$$L(x, z; \mu) = \sum_{r \in R} \frac{\sqrt{2}}{T_r} \arctan\left(\frac{x_r T_r}{\sqrt{2}}\right) + \mu \cdot (C - Ax - z),$$

where $\mu = (\mu_j, j \in J)$ is a vector of Lagrange multipliers, and $z = C - Ax$ is a vector of *slack variables*, measuring the spare capacity on each of the links $j \in J$ of the network. Then, using the derivative of the arctan function,

$$\frac{\partial L}{\partial x_r} = (1 + \frac{1}{2} x_r^2 T_r^2)^{-1} - \sum_{j \in J} \mu_j A_{jr} \quad \text{and} \quad \frac{\partial L}{\partial z_j} = -\mu_j.$$

We look for a maximum of L over $x, z \geq 0$; it turns out that this maximum is, under the identification $\mu_j = p_j$, precisely the collection $(x_r, r \in R)$, $(p_j, j \in J)$ of rates and drop probabilities that we were looking for. For example, setting to zero the partial derivative with respect to x_r gives the desired equation for x_r .

In summary, for each link $j \in J$ the Lagrange multiplier μ_j arising from the optimization problem is pre-

cisely the proportion p_j of packets dropped at that link, much as the Lagrange multipliers arising earlier were precisely the delays on links of a road traffic network. And the equilibrium reached by the interaction of many competing TCPs, each implemented only on the source and destination computers, is effectively maximizing an objective function for the entire network. The objective function has a surprising interpretation: it is as if the usefulness of the flow rate x_r to the source-destination pair served by this route is given by a *utility function*

$$\frac{\sqrt{2}}{T_r} \arctan\left(\frac{x_r T_r}{\sqrt{2}}\right),$$

and the network is attempting to maximize the sum of these utility functions across all source-destination pairs, subject to constraints arising from the limited capacities of the links.

The arctan function, illustrated in figure 4, is *concave*. Thus, if two or more connections share an overloaded link, the rates achieved will be approximately equal, since otherwise the total utility could be increased by reducing the largest rate a little and increasing the smallest rate a little. As a result, there is a tendency for TCP to share resources more or less equitably. This is very different from resource-control mechanisms in traditional telephone networks where, if the network is overloaded, some calls are blocked in order that the calls that are accepted are unaffected by the overload.

6 Conclusion

The behavior of large-scale systems has been of great interest to mathematicians for over a century, with many examples coming from physics. For example, the behavior of a gas can be described at the microscopic level in terms of the position and velocity of each molecule. At this level of detail a molecule's velocity appears as a random process, as the molecule bounces around off other molecules and the walls of the container. Yet consistent with this detailed microscopic description of the system is macroscopic behavior best described by quantities such as temperature and pressure. Similarly, the behavior of electrons in an electrical network can be described in terms of random walks, and yet this simple description at the microscopic level leads to rather sophisticated behavior at the macroscopic level: Kelvin showed that the pattern of potentials in a network of resistors is exactly the one that minimizes heat dissipation for a given level of current flow (Kelly 1991). The local, random behavior

of the electrons causes the network as a whole to solve a rather complex optimization problem.

In the last fifty years we have begun to realize that large-scale engineered systems are often best understood in similar terms. Thus a microscopic description of traffic flow in terms of each driver's choice of the most convenient route can be consistent with macroscopic behavior described in terms of a function minimization. And the simple, local rules that control how packets are transmitted through the Internet can correspond to a maximizing of aggregate utility across the entire network.

One thought-provoking difference is that, whereas the microscopic rules governing physical systems are fixed, for engineered systems such as transport or communication networks we may be able to choose the microscopic rules so as to achieve the macroscopic consequences we judge desirable.

Further Reading

- Beckmann, M., C. B. McGuire, and C. B. Winsten. 1956. *Studies in the Economics of Transportation*. Cowles Commission Monograph. New Haven, CT: Yale University Press.
- Braess, D. 1968. Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung* 12:258–68.
- Cohen, J. E. 1988. The counterintuitive in conflict and cooperation. *American Scientist* 76:576–84.
- Department for Transport. 2004. Feasibility study of road pricing in the UK. Available from www.dft.gov.uk.
- Jacobson, V. 1988. Congestion avoidance and control. *Computer Communication Review* 18(4):314–29.
- Kelly, F. P. 1991. Network routing. *Philosophical Transactions of the Royal Society of London A* 337:343–67.
- . 2001. Mathematical modeling of the Internet. In *Mathematics Unlimited—2001 and Beyond*, edited by B. Engquist and W. Schmid, pp. 685–702. Berlin: Springer.
- Low, S. H., F. Paganini, and J. C. Doyle. 2002. Internet congestion control. *IEEE Control Systems Magazine* 22: 28–43.
- Wardrop, J. G. 1952. Some theoretical aspects of road traffic research. *Proceedings of the Institute of Civil Engineers* 1: 325–78.

VII.5 The Mathematics of Algorithm Design

Jon Kleinberg

1 The Goals of Algorithm Design

When computer science began to emerge as a subject at universities in the 1960s and 1970s, it drew some

amount of puzzlement from the practitioners of more established fields. Indeed, it is not initially clear why computer science should be viewed as a distinct academic discipline. The world abounds with novel technologies, but we do not generally create a separate field around each one; rather, we tend to view them as by-products of existing branches of science and engineering. What is special about computers?

Viewed in retrospect, such debates highlighted an important issue: computer science is not so much about the computer as a specific piece of technology as it is about the more general phenomenon of computation itself, the design of processes that represent and manipulate information. Such processes turn out to obey their own inherent laws, and they are performed not only by computers but by people, by organizations, and by systems that arise in nature. We will refer to these computational processes as *algorithms*. For the purposes of our discussion in this article, one can think of an algorithm informally as a step-by-step sequence of instructions, expressed in a stylized language, for solving a problem.

This view of algorithms is general enough to capture both the way a computer processes data and the way a person performs calculations by hand. For example, the rules for adding and multiplying numbers that we learn as children are algorithms; the rules used by an airline company for scheduling flights constitute an algorithm; and the rules used by a search engine like Google for ranking Web pages constitute an algorithm. It is also fair to say that the rules used by the human brain to identify objects in the visual field constitute a kind of algorithm, though we are currently a long way from understanding what this algorithm looks like or how it is implemented on our neural hardware.

A common theme here is that one can reason about all these algorithms without recourse to specific computing devices or computer programming languages, instead expressing them using the language of mathematics. In fact, the notion of an algorithm as we now think of it was formalized in large part by the work of mathematical logicians in the 1930s, and algorithmic reasoning is implicit in the past several millennia of mathematical activity. (For example, equation-solving methods have always tended to have a strong algorithmic flavor; the geometric constructions of the ancient Greeks were inherently algorithmic as well.) Today, the mathematical analysis of algorithms occupies a central position in computer science; reasoning about algorithms independently of the specific devices on which

they run can yield insight into general design principles and fundamental constraints on computation.

At the same time, computer-science research struggles to keep two diverging views in focus: this more abstract view that formulates algorithms mathematically, and the more applied view that the public generally associates with the field, the one that seeks to develop applications such as Internet search engines, electronic banking systems, medical imaging software, and the host of other creations we have come to expect from computer technology. The tension between these two views means that the field's mathematical formulations are continually being tested against their implementation in practice; it provides novel avenues for mathematical notions to influence widely used applications; and it sometimes leads to new mathematical problems motivated by these applications.

The goal of this short article is to illustrate this balance between the mathematical formalism and the motivating applications of computing. We begin by building up to one of the most basic definitional questions in this vein: how should we formulate the notion of *efficient* computation?

2 Two Representative Problems

To make the discussion of efficiency more concrete, and to illustrate how one might think about an issue like this, we first discuss two representative problems—both fundamental in the study of algorithms—that are similar in their formulation but very different in their computational difficulty.

The first in this pair is the *traveling-salesman problem* (TSP), which is defined as follows. We imagine a salesman contemplating a map with n cities (he is currently located in one of them). The map gives the distance between each pair of cities, and the salesman wishes to plan the shortest possible tour that visits all n cities and returns to the starting point. In other words, we are seeking an algorithm that takes as input the set of all distances among pairs of cities, and produces a tour of minimum total length. Figure 1(a) depicts the optimal solution to a sample input instance of the TSP; the circles represent the cities, the dark lines (with lengths labeling them) connect cities that the salesman visits consecutively on the tour, and the lighter lines connect all the other pairs of cities, which are not visited consecutively.

A second problem is the *minimum-spanning-tree problem* (MSTP). Here we imagine a construction firm

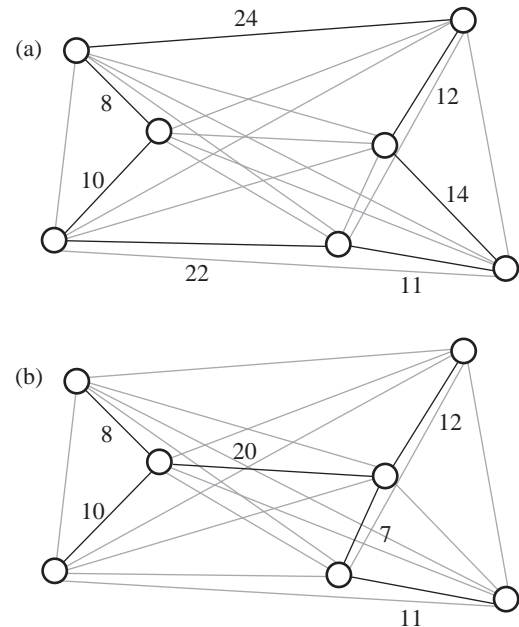


Figure 1 Solutions to instance of (a) the traveling-salesman problem and (b) the minimum-spanning-tree problem, on the same set of points. The dark lines indicate the pairs of cities that are connected by the respective optimal solutions, and the lighter lines indicate all pairs that are not connected.

with access to the same map of n cities, but with a different goal in mind. They wish to build a set of roads connecting certain pairs of the cities on the map, so that after these roads are built there is a route from each of the n cities to each other one. (A key point here is that each road must go directly from one city to another.) The goal is to build such a road network as cheaply as possible—in other words, using as little total road material as possible. Figure 1(b) depicts the optimal solution to the instance of the MSTP defined by the same set of cities used for part (a).

Both of these problems have a wide range of practical applications. The TSP is a basic problem concerned with sequencing a given set of objects in a “good” order; it has been used for problems that run from planning the motion of robotic arms drilling holes on printed circuit boards (where the “cities” are the locations where the holes must be drilled) to ordering genetic markers on a chromosome in a linear sequence (with the markers constituting the cities, and the distances derived from probabilistic estimates of proximity). The MSTP is a basic issue in the design of efficient communica-

Terri: Tim prefers this as it is. OK?

tion networks; this follows the motivation given above, with fiber-optic cable acting in the role of “roads.” The MSTP also plays an important role in the problem of clustering data into natural groupings. Note, for example, how the points on the left-hand side of figure 1(b) are joined to the points on the right-hand side by a relatively long link; in clustering applications, this can be taken as evidence that the left and right points form natural groupings.

It is not hard to come up with an algorithm for solving the TSP. We first list every possible way of ordering the cities (other than the starting city, which is fixed in advance). Each ordering defines a tour—the salesman could visit the cities in this order and then return to the start—and for each ordering we could compute the total length of the tour, by traversing the cities in this order and summing the distances from each city to the next. As we perform this calculation for all possible orders, we keep track of the order that yields the smallest total distance, and at the end of the process we return this tour as the optimal solution.

While this algorithm does solve the problem, it is extremely inefficient. There are $n - 1$ cities other than the starting point, and any possible sequence of them defines a tour, so we need to consider $(n - 1)(n - 2)(n - 3) \cdots (3)(2)(1) = (n - 1)!$ possible tours. Even for $n = 30$ cities, this is an astronomically large quantity; on the fastest computers we have today, running this algorithm to completion would take longer than the life expectancy of the Earth. The difficulty is that the algorithm we have just described is performing a *brute-force search*: the “search space” of possible solutions to the TSP is very large, and the algorithm is doing nothing more than plowing its way through this entire space, considering every possible solution.

For most problems, there is a comparably inefficient algorithm that simply performs a brute-force search. Things tend to get interesting when one finds a way to improve significantly on this brute-force approach.

The MSTP provides a nice example of how such an improvement can happen. Rather than considering all possible road networks on the given set of cities, suppose we try the following myopic, “greedy” approach to the MSTP. We sort all the pairs of cities in order of increasing distance, and then work through the pairs in this order. When we get to a pair of cities, say A and B, we test if there is already a way to travel from A to B in the collection of roads constructed thus far. If there is, then it would be superfluous to build a direct road from A to B—our goal, remember, is just to make

sure every pair is connected by *some* sequence of roads, and A and B are already connected in this case. But if there is no way to get from A to B using what has already been built, then we construct the direct road from A to B. (As an example of this reasoning, note that the potential road of length 14 in figure 1(a) would not get built by this MSTP algorithm; by the time this direct route is considered, its endpoints are already joined by the sequence of two shorter roads of length 7 and 11, as depicted in figure 1(b).)

It is not at all obvious that the resulting road network should have the minimum possible cost, but in fact this is true. In other words, one can prove a theorem that says, essentially, “On every input, the algorithm just described produces an optimal solution.” The payoff from this theorem is that we now have a way to compute an optimal road network by an algorithm that is much, much more efficient than brute-force search: it simply needs to sort the pairs of cities by their distances, and then make a single pass through this sorted list to decide which roads to build.

This discussion has provided us with a fair amount of insight into the nature of the TSP and the MSTP. Rather than experimenting with actual computer programs, we described algorithms in words, and made claims about their performance that could be stated and proved as mathematical theorems. But what can we abstract from these examples if we want to talk about computational efficiency in general?

3 Computational Efficiency

Most interesting computational problems share the following feature with the TSP and the MSTP: an input of size n implicitly defines a search space of possible solutions whose size grows exponentially with n . One can appreciate this explosive growth rate as follows: if we simply add one to the size of the input, the time required to search the entire space increases by a multiplicative factor. We would prefer algorithms to scale more reasonably: their running times should only increase by a multiplicative factor when the input itself increases by a multiplicative factor. Running times that are bounded by a polynomial function of the input size—in other words, proportional to n raised to some fixed power—exhibit this property. For example, if an algorithm requires at most n^2 steps on an input of size n , then it requires at most $(2n)^2 = 4n^2$ steps on an input twice as large.

In part because of arguments like this, computer scientists in the 1960s adopted *polynomial time* as a working definition of efficiency: an algorithm is deemed to be efficient if the number of steps it requires on an input of size n grows like n raised to a fixed power. Using the concrete notion of polynomial time as a surrogate for the fuzzier concept of efficiency is the kind of modeling decision that ultimately succeeds or fails based on its utility in guiding the development of real algorithms. And in this regard, polynomial time has turned out to be a definition of surprising power in practice: problems for which one can develop a polynomial-time algorithm have turned out in general to be highly tractable, while those for which we lack polynomial-time algorithms tend to pose serious challenges even for modest input sizes.

A concrete mathematical formulation of efficiency provides a further benefit: it becomes possible to pose, in a precise way, the conjecture that certain problems cannot be solved by efficient algorithms. The TSP is a natural candidate for such a conjecture; after decades of failed attempts to find an efficient algorithm for the TSP, one would like to be able to prove a theorem that says, “There is no polynomial-time algorithm that finds an optimal solution to every instance of the TSP.” A theory known as NP-COMPLETENESS [IV.20 §4] provides a unifying framework for thinking about such questions; it shows that a large class of computational problems, containing literally thousands of naturally arising problems (including the TSP), are equivalent with respect to polynomial-time solvability: there is an efficient algorithm for one if and only if there is an efficient algorithm for all. It is a major open problem to decide whether or not these problems have efficient algorithms; the deeply held sense that they do not has become the *P versus NP conjecture*, which has begun to appear on lists of the most prominent problems in mathematics.

Like any attempt to make an intuitive notion mathematically precise, polynomial time as a definition of efficiency in practice begins to break down around its boundaries. There are algorithms for which one can prove a polynomial bound on the running time, but which are hopelessly inefficient in practice. Conversely, there are well-known algorithms (such as the standard SIMPLEX METHOD [III.86] for linear programming) that require exponential running time on certain pathological instances, but which run quickly on almost all inputs encountered in real life. And for computing applications that work with massive data sets, an algo-

rithm with a polynomial running time may not be efficient enough; if the input is a trillion bytes long (as can easily occur when dealing with snapshots of the Web, for example), even an algorithm whose running time depends quadratically on the input will be unusable in practice. For such applications, one generally needs algorithms that scale linearly with the size of the input—or, more strongly, that operate by “streaming” through the input in one or two passes, solving the problem as they go. The theory of such streaming algorithms is an active topic of research, drawing on techniques from information theory, Fourier analysis, and other areas. None of this means that polynomial time is losing its relevance to algorithm design—it is still the standard benchmark for efficiency—but new computing applications tend to push the limits of current definitions, and in the process raise new mathematical problems.

4 Algorithms for Computationally Intractable Problems

In the previous section we discussed how researchers have identified a large class of natural problems, including the TSP, for which it is strongly believed that no efficient algorithm exists. While this explains our difficulties in solving these problems optimally, it leaves open a natural question: what should we do when actually confronted by such a problem in practice?

There are a number of different strategies for approaching such computationally intractable problems. One of these is *approximation*: for problems like the TSP that involve choosing an optimal solution from among many possibilities, we could try to formulate an efficient algorithm that is guaranteed to produce a solution almost as good as the optimal one. The design of such approximation algorithms is an active area of research; we can see a basic example of this process by considering the TSP. Suppose we are given an instance of the TSP, specified by a map with distances, and we set ourselves the task of constructing a tour whose total length is at most twice that of the shortest tour. At first this goal seems a bit daunting: since we do not know how to compute the optimal tour (or its length), how will we guarantee that the solution we produce is short enough? It turns out, however, that this can be done by exploiting an interesting connection between the TSP and the MSTP, a relationship between the respective optimal solutions to each problem on the same set of cities.

Consider an optimal solution to the MSTP on the given set of cities, consisting of a network of roads; recall that this is something we can compute efficiently. Now, the salesman interested in finding a short tour for these cities can use this optimal road network to visit the cities as follows. Starting at one city, he follows roads until he hits a dead end, that is, a city with no new roads exiting it. He then backs up, retracing his steps until he gets to a junction with a road he has not yet taken, and he proceeds down this new road. For example, starting in the upper left corner of figure 1(b), the salesman would follow the road of length 8 and then choose one of the roads of length 10 or 20; if he selects the former, then after reaching the dead end he would back up to this junction again and continue the tour by following the road of length 20. A tour constructed in this way traverses each road twice (once in each direction), so if we let m denote the total length of all roads in the optimal MSTP solution, we have found a tour of length $2m$.

How does this compare to t , the length of the best possible tour? Let us first argue that $t \geq m$. This is true because, in the space of all possible solutions to the MSTP, one option is to build roads between cities that the salesman visits consecutively in the optimal TSP tour, for a total mileage of t ; on the other hand, m is the total length of the *shortest possible* road network, and hence t cannot be smaller than m . So we have concluded that the optimal solution to the TSP has length at least m . However, we have just exhibited an algorithm that finds a tour of length $2m$, so, as we wanted, we have an efficient way to find a tour that is at most twice as long as the shortest one possible.

People trying to solve large instances of computationally hard problems in practice frequently use algorithms that have been observed empirically to give nearly optimal solutions, even when no guarantees on their performance have been proved. *Local-search* algorithms form one widely used class of approaches like this. A local-search algorithm starts with an initial solution and repeatedly modifies it by making some “local” change to its structure, looking for a way to improve its quality. In the case of the TSP, a local-search algorithm would seek simple improving modifications to its current tour; for example, it might look at sets of cities that are visited consecutively and see if visiting them in the opposite order would shorten the tour. Researchers have drawn connections between local-search algorithms and phenomena in nature; for example, just as a large molecule contorts itself in space

trying to find a minimum-energy conformation, we can imagine the TSP tour in a local-search algorithm modifying itself as it tries to reduce its length. Determining how deeply this analogy goes is an interesting research issue.

5 Mathematics and Algorithm Design: Reciprocal Influences

Many branches of mathematics have contributed to aspects of algorithm design, and the issues raised by the analysis of new algorithmic problems have, in a number of cases, suggested novel mathematical questions.

Combinatorics and graph theory have been qualitatively transformed by the growth of computer science, to the extent that algorithmic questions have become thoroughly intertwined with the mainstream of research in these areas. Techniques from probability have also become fundamental to many areas of computer science: probabilistic algorithms draw power from the ability to make random choices while they are being executed, and probabilistic models of the input to an algorithm can give one a more realistic view of the problem instances that arise in practice. This style of analysis provides a steady source of new questions in discrete probability.

A computational perspective is often useful in thinking about “characterization” problems in mathematics. For example, the general issue of characterizing prime numbers has an obvious algorithmic component: given a number n as input, how efficiently can we determine whether it is prime? (There exist algorithms that are exponentially better than the approach of dividing n by all numbers up to \sqrt{n} : see COMPUTATIONAL NUMBER THEORY [IV.3 §2].) Problems in KNOT THEORY [III.46], such as the characterization of unknotted loops, have a similar algorithmic side. Suppose we are given a circular loop of string in three dimensions (described as a jointed chain of line segments), and it wraps around itself in complicated ways. How efficiently can we determine whether it is truly knotted, or whether by moving it around we can fully untangle it? We can ask this sort of question in many similar mathematical contexts; it is clear that these algorithmic issues are extremely concrete as problems, though they may lose part of the original intent of the mathematicians who posed the questions more generally.

Rather than attempting to enumerate the intersection of algorithmic ideas with all the different branches

of mathematics, we conclude this article with two case studies that involve the design of algorithms for particular applications, and the ways in which mathematical ideas arise in each instance.

6 Web Search and Eigenvectors

As the World Wide Web grew in popularity throughout the 1990s, computer-science researchers grappled with a difficult problem: the Web contains a vast amount of useful information, but its anarchic structure makes it very hard for users, unassisted, to find the specific information they are looking for. Thus, early in the Web's history, people began to develop *search engines* that would index the information on the Web, and produce relevant Web pages in response to user queries. But of the thousands or millions of pages relevant to a topic on the Web, which few should the search engine present to a user? This is the *ranking* problem: how to determine the “best” resources on a given topic. Note the contrast with concrete problems like the TSP. There, the goal (the shortest tour) was not in doubt; the difficulty was simply in computing an optimal solution efficiently. For the search engine ranking problem, on the other hand, formalizing the goal is a large part of the challenge—what do we mean by the “best” page on a topic? In other words, an algorithm to rank Web pages is really providing a *definition* of the quality of a Web page as well as the means to evaluate this definition.

The first search engines ranked each Web page based purely on the text it contained. These approaches began to break down as the Web grew, because they did not take into account the quality judgments encoded in the Web's hyperlinks: in browsing the Web, we often discover high-quality resources because they are “endorsed” through the links they receive from other pages. This insight led to a second generation of search engines that determined rankings using *link analysis*.

The simplest such analysis would just count the number of links to a page: in response to the query “newspapers,” for example, one could rank pages by the number of incoming links they receive from other pages containing the term—in effect, allowing pages containing the term “newspapers” to vote on the result. Such a scheme will generally do well for the top few items, placing prominent news sites like *The New York Times* and *The Financial Times* at the head of the list; beyond this, however, it will quickly break down, favoring a large number of highly linked but irrelevant sites.

It is possible to make much more effective use of the latent information in the links. Consider pages that link

to many of the sites ranked highly by this simple voting scheme; it is natural to expect that these are authored by people with a good sense for where the interesting newspapers are, and so we could run the voting again, this time giving more voting power to these pages that selected many of the highly ranked sites. This revote might elevate certain lesser-known newspapers favored by Web-page authors who were more knowledgeable on the topic; in response to the results of this revote, we could further sharpen our weighting of the voters. This “principle of repeated improvement” uses the information contained in a set of page-quality estimates to produce a more refined set of estimates. If we perform these refinements repeatedly, will they converge to a stable solution?

In fact, this sequence of refinements can be viewed as an algorithm for computing the principal EIGENVECTOR [I.3 §4.3] of a particular matrix (see SOME FUNDAMENTAL MATHEMATICAL DEFINITIONS [I.3]); this both establishes the convergence of the process and characterizes the end result. To establish this connection, we introduce some notation. Each Web page is assigned two scores: an *authority weight*, measuring its quality as a primary source on the topic; and a *hub weight*, measuring its power as a voter for the highest-quality content. Pages may score highly in one of these measures but not in the other—one should not expect a prominent newspaper to simultaneously serve as a good guide to other newspapers—but there is also nothing to prevent a page from scoring well in both. One round of voting can now be viewed as follows: we update the authority weight of each page by summing the hub weights of all pages that point to it (receiving links from highly weighted voters makes you a better authority); we then reweight all the voters, updating each page's hub weight by summing the authority weights of the pages it points to (linking to high-quality content makes you a better hub).

How do eigenvectors come into this? Suppose we define a matrix M with one row and one column for each page under consideration; the (i, j) entry equals 1 if page i links to page j , and it equals 0 otherwise. We encode the authority weights in a vector \mathbf{a} , where the coordinate a_i is the authority weight of page i . The hub weights can be similarly written as a vector \mathbf{h} . Using the definition of matrix-vector multiplication, we can now check that the updating of hub weights in terms of authority weights is simply the act of setting \mathbf{h} equal to $M\mathbf{a}$; correspondingly, setting \mathbf{a} equal to $M^T\mathbf{h}$ updates the authority weights. (Here M^T denotes

the transpose of the matrix M .) Running these updates n times each from starting vectors a_0 and h_0 , we obtain $\mathbf{a} = (M^T(M(M^T(M \cdots (M^T(Ma_0)) \cdots)))) = (M^T M)^n a_0$. This is the power-iteration method for computing the principal eigenvector of $M^T M$, in which we repeatedly multiply some fixed starting vector by larger and larger powers of $M^T M$. (As we do this, we also divide all coordinates of the vector by a scaling factor to prevent them from growing unboundedly.) Hence this eigenvector is the stable set of authority weights toward which our updates are converging. By completely symmetric reasoning, the hub weights are converging toward the principal eigenvector of MM^T .

A related link-based measure is *PageRank*, defined by a different procedure that is also based on repeated refinement. Instead of drawing a distinction between the voters and the voted-on, one posits a single kind of quality measure that assigns a *weight* to each page. A current set of page weights is then updated by having each page distribute its weight uniformly among the pages it links to. In other words, receiving links from high-quality pages raises one's own quality. This too can be written as multiplication by a matrix, obtained from M^T by dividing each row's entries by the number of outgoing links from the corresponding page; repeated updates again converge to an eigenvector. (There is a further wrinkle here: repeated updating in this case tends to cause all weight to pool at "dead-end" pages that have no outgoing links and hence nowhere to pass their weight. Thus, to obtain the PageRank measure used in applications, one adds a tiny quantity $\varepsilon > 0$ in each iteration to the weight of each page; this is equivalent to using a slightly modified matrix.)

PageRank is one of the main ingredients in the search engine Google; hubs and authorities form the basis for Ask's search engine Teoma, as well as a number of other Web search tools. In practice, current search engines (including Google and Ask) use highly refined versions of these basic measures, often combining features of each; understanding how relevance and quality measures are related to large-scale eigenvector computations remains an active research topic.

7 Distributed Algorithms

Thus far we have been discussing algorithms that run on a single computer. As a concluding topic, we briefly touch on a broad area in computer science concerned with computations that are *distributed* over multiple communicating computers. Here the problem of efficiency is compounded by concerns over maintaining

coordination and consistency among the communicating processes.

As a simple example illustrating these issues, consider a network of automatic teller machines (ATMs). When you withdraw an amount of money x at one of these ATMs, it must do two things: (1) notify a central bank computer to deduct x from your account; and (2) emit the correct amount of money in physical bills. Now, suppose that between steps (1) and (2) the ATM crashes so that you do not get your money; you would like it to be the case that the bank does not subtract x from your account anyway. Or suppose that the ATM executes both of steps (1) and (2), but its message to the bank is lost; the bank would like for x to be eventually subtracted from your account anyway. The field of distributed computing is concerned with designing algorithms that operate correctly in the presence of such difficulties.

As a distributed system runs, certain processes may experience long delays, some of them may fail in mid-computation, and some of the messages between them may be lost. This leads to significant challenges in reasoning about distributed systems, because this pattern of failures can cause each process to have a slightly different *view* of the computation. It is easily possible for there to be two runs of the system, with different patterns of failure, that are "indistinguishable" from the point of view of some process P ; in other words, P will have the same view of each, simply because the differences in the runs did not affect any of the communications that it received. This can pose a problem if P 's final output is supposed to depend on its having noticed that the two runs were different.

A major advance in the study of such systems came about in the 1990s, when a connection was made to techniques from algebraic topology. Consider for simplicity a system with three processes, though everything we say generalizes to any number of processes. We consider the set of all possible runs of the system; each run defines a set of three views, one held by each process. We now imagine the views associated with a single run as the three corners of a triangle, and we glue these triangles together according to the following rule: for any two runs that are indistinguishable to some process P , we paste the two corresponding triangles together at their corners associated with P . This gives us a potentially very complicated geometric object, constructed by applying all these pasting operations to the triangles; we call this object the *complex* associated with the algorithm. (If there were more than

three processes, we would have an object in a higher number of dimensions.) While it is far from obvious, researchers have been able to show that the correctness of distributed algorithms can be closely connected with the topological properties of the complexes that they define.

This is another powerful example of the way in which mathematical ideas can appear unexpectedly in the study of algorithms, and it has led to new insights into the limits of the distributed model of computation. Combining the analysis of algorithms and their complexes with classical results from algebraic topology has in some cases resolved tricky open problems in this area, establishing that certain tasks are provably impossible to solve in a distributed system.

Further Reading

Algorithm design is a standard topic in the undergraduate computer-science curriculum, and it is the subject of a number of textbooks, including Cormen et al. (2001) and a book by Kleinberg and Tardos (2005). The perspective of early computer scientists on how to formalize efficiency is discussed by Sipser (1992). The TSP and the MSTP are fundamental to the field of combinatorial optimization; the TSP is used as a lens through which this field is surveyed in a book edited by Lawler et al. (1985). Approximation algorithms and local-search algorithms for computationally intractable problems are discussed in books edited by Hochbaum (1996) and by Aarts and Lenstra (1997), respectively. Web search and the role of link analysis is covered in a book by Chakrabarti (2002); beyond Web applications, there are a number of other interesting connections between eigenvectors and network structures, as described by Chung (1997). Distributed algorithms are covered in a book by Lynch (1996), and the topological approach to analyzing distributed algorithms is reviewed by Rajsbaum (2004).

- Aarts, E., and J. K. Lenstra, eds. 1997. *Local Search in Combinatorial Optimization*. New York: John Wiley.
- Chakrabarti, S. 2002. *Mining the Web*. San Mateo, CA: Morgan Kaufman.
- Chung, F.R.K. 1997. *Spectral Graph Theory*. Providence, RI: American Mathematical Society.
- Cormen, T., C. Leiserson, R. Rivest, and C. Stein. 2001. *Introduction to Algorithms*. Cambridge, MA: MIT Press.
- Hochbaum, D. S., ed. 1996. *Approximation Algorithms for NP-hard Problems*. Boston, MA: PWS Publishing.
- Kleinberg, J., and É. Tardos. 2005. *Algorithm Design*. Boston, MA: Addison-Wesley.

- Lawler, E. L., J. K. Lenstra, A.H.G. Rinnooy Kan, and D. B. Shmoys, eds. 1985. *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*. New York: John Wiley.
- Lynch, N. 1996. *Distributed Algorithms*. San Mateo, CA: Morgan Kaufman.
- Rajsbaum, S. 2004. Distributed computing column 15. *ACM SIGACT News* 35:3.
- Sipser, M. 1992. The history and status of the P versus NP question. In *Proceedings of the 24th ACM Symposium on Theory of Computing*. New York: Association for Computing Machinery.

VII.6 Reliable Transmission of Information

Madhu Sudan

1 Introduction

The notion of “digital information” emerged in the middle of the twentieth century, in response to the advent of the telegraph and to the beginnings of computer science, which at the time was principally a theoretical discipline. Of course, the use of electricity to communicate signals goes back further, but the earlier uses involved signals of a “continuous” nature: music, voice, etc. The new era was characterized by the transmission of (or the need to transmit) more “discrete” messages, i.e., messages such as English sentences, which can be described as finite sequences of letters taken from some finite alphabet. The phrase “digital information” came to be applied to such families of messages.

Digital information posed some novel challenges to the engineers and mathematicians charged with the task of communicating such messages. The root cause of these challenges is “noise.” Every communication medium is noisy, and never transmits any signal completely accurately. In the case of continuous signals, somehow the receivers (typically, our ears and eyes) can adjust to such errors and learn to discount them. For example, if you play a very old recording of a musical performance, then there will typically be a crackling noise, but it is possible to ignore this, unless the quality is very bad indeed, and concentrate on the music. However, in the case of digital information errors can have a more catastrophic effect. To see this, suppose that we are communicating in English sentences and that the communication medium makes occasional mistakes by altering one of the transmitted letters. In such a scenario the message

WE ARE NOT READY

could easily be changed into the message

WE ARE NOW READY.

All it takes is one error on the part of the communication medium, and the entire intention of the message is reversed. Digital information tends to be inherently intolerant of errors, and the mathematicians and engineers of the time were charged with the task of inventing methods that would make communication reliable even if the process of transmission is not.

Here is one way of achieving this. To communicate any message, the sender of the message repeats every letter, say five times. For example, to send the message

WE ARE NOT READY

the sender says something like

WWWWWEIEEE AAAAA...

The receiver can then detect errors (as long as there are not too many) by checking that every block of five successive letters repeats the same letter. If this ever fails to be the case, then it is clear that errors have occurred during transmission. If it is not possible for five successive symbols to be in error (or even if it is just very unlikely), then it follows that the resulting scheme is also more reliable than the underlying means of transmission. Finally, if even less error is possible, then it may be possible for the receiver to determine the actual message, rather than simply being able to tell when errors have occurred. For example, if at most two symbols in any block of five can be erroneous, then the most commonly occurring letter in each block of five must be the letter from the original message: for instance, a sequence such as

WWWMWEFEEE AAAAA...

would be interpreted by the receiver as

WE A...

Repeating every symbol five times in order to be able to correct two errors does not appear to be a very efficient way to use the communication channel. Indeed, as we will show in the rest of this article, when transmitting long messages one can do much better. However, in order to understand this issue, we need to define the process of communication, the model of error, and the measures of performance more carefully. We do so next.

2 Model

2.1 Channel and Errors

The central object of attention in the problem of information transmission is the “channel of communication,” or simply the *channel*. The channel has an *input* (the original signal to be communicated) and an *output* (the signal after it is transmitted). The input consists of a sequence of elements from some finite set: by analogy with the English-language example, these elements are called *letters* and the finite set, which is typically denoted Σ , is called an *alphabet*. The channel attempts to transmit the input to the receiver, but while doing so it may make some errors. The alphabet and the process that underlies the errors are what specifies the channel.

The alphabet Σ varies from scenario to scenario. In the example described above, the alphabet consisted of the English characters $\{A, B, \dots, Z\}$, and possibly some punctuation symbols. In most communication scenarios, the alphabet is the “binary alphabet” that consists just of the “letters” 0 and 1, which are known as *bits*. On the other hand, in applications involving storage of digital information (in compact discs (CDs), digital versatile discs (DVDs), etc.), the alphabet contains 256 elements (the alphabet of “bytes”).

Specifying an alphabet is easy, but if we wish to define a good mathematical model for the way that errors are produced, then a lot more care is needed. At one extreme is a worst-case model suggested by Hamming (1950), where there is some limit on the number of errors that the channel can make, but within that limit it chooses the errors to be as damaging as possible. A more benign class of errors was proposed by Shannon (1948), who suggested that errors could be modeled by a probabilistic process.

We will focus on one probabilistic model to illustrate many of the concepts below. In this model, the error of the channel is specified by a real number parameter p , where $0 \leq p \leq 1$. Every use of the channel results in an error with probability p . To be precise, if the sender transmits an element $\sigma \in \Sigma$, then with probability $1 - p$ the output for that element is σ but with probability p it is some other element σ' of Σ , chosen uniformly at random. Furthermore, and this is very crucial to this model, the errors are assumed to be *independent*, i.e., the channel repeats this process for each letter it transmits without any memory of how it acted on previous symbols. We refer to this model as the Σ -symmetric channel with parameter p (or Σ -SC(p)) in the rest of this article. A special case of particular

Terri: this is fine according to Tim.
OK?

importance is the *binary symmetric channel*, which is the Σ -symmetric channel when Σ is the binary alphabet $\{0, 1\}$. Then, if the input bit is 0, say, the corresponding output bit will be 0 with probability $1 - p$ and 1 with probability p .

While this model of error may seem rather oversimplified (and even unnatural if Σ is not the binary alphabet $\{0, 1\}$), it turns out that it captures the essence of most mathematical challenges that arise when one tries to make communication reliable. Furthermore, many of the solutions found to make communication reliable in this setting have been generalized to other scenarios, so this simple model is very useful both in practice and in the theoretical study of communication.

2.2 Encoding and Decoding

Suppose the sender wishes to transmit a sequence through a channel that makes errors. One way to compensate for these errors is to send through the channel not the sequence itself but a modified version of the sequence that contains redundant information. The process of modification that we choose is called the *encoding* of the message. We have already seen one method of encoding, namely repeating each term in the sequence several times. However, this is by no means the only way of doing it, so to discuss encoding we use the following general framework: if the sender has a message consisting of a sequence of k elements of Σ , then by some means or another it expands the message into a new sequence, now consisting of n elements of Σ , for some $n > k$. Formally, the sender applies an *encoding function* $E : \Sigma^k \rightarrow \Sigma^n$ to the message. (Σ^k stands for the set of sequences of length k with letters in Σ , and Σ^n for the set of sequences of length n .) Thus, to convey a message $m = (m_1, m_2, \dots, m_k)$ to the receiver, the sender transmits over the channel not the k symbols of m but the n symbols of $E(m)$.

The receiver now receives a sequence $r = (r_1, r_2, \dots, r_n)$, belonging to Σ^n , and its goal is to “compress” this sequence back to a k -letter sequence, removing the error and obtaining the original message m (at least if not too many errors have occurred). It does this by applying a *decoding function* $D : \Sigma^n \rightarrow \Sigma^k$, which tells it how sequences of length n are converted back into sequences of length k .

The possible pairs of functions E, D describe the options available to the designers of the communication system. Their choice determines the performance of the system. Let us now describe how this performance is measured.

2.3 Goals

Very informally, our goals are threefold. We would like to make the communication as reliable as possible. At the same time, we would like to maximize the utilization of the channel. Finally, we would like to do so with effective computation. We describe these goals more carefully below, in the case of the model $\Sigma\text{-SC}(p)$ described earlier.

Consider first the reliability. If we start with a message m , encode it as $E(m)$, and pass it through the channel, then the output, after some random errors have been introduced, will be a string y . The receiver will decode y , producing a new message $D(y)$. For each message m , there is a certain probability of a *decoding error*, i.e., a certain probability that $D(y)$ will not in fact be equal to the original message m . The reliability of the communication is measured by the largest of these probabilities. If this is small, then we know that, whatever the original message m , a decoding error is unlikely, and then we regard the communication as reliable.

Next, let us look at the utilization of the channel. This is measured by the *rate* of the encoding, i.e., the quantity k/n . In other words, it is the ratio of the length of the original message to the length of the encoded message: the smaller this ratio, the less efficiently one is using the channel.

Finally, practical considerations also require us to be able to encode and decode quickly: a pair of reliable and efficient encoding and decoding functions will not be of much use if they are very time-consuming to compute. Adopting the standard convention in algorithm design, we regard our algorithms as feasible if they run in *polynomial time*: that is, if their running time can be bounded above by a polynomial function of the length of their input and output.

To illustrate the above ideas, let us analyze the “repetition encoding” that repeats every letter of the alphabet five times. For simplicity, take the alphabet Σ to be $\{0, 1\}$, let the probability p be fixed, and let us consider the behavior of the model as the message length k tends to ∞ . Our encoding function takes strings of length k to strings of length $5k$ and thus has a rate of $\frac{1}{5}$. Given any particular block of five transmissions, the probability that it contains three or more errors is

$$p' = \binom{5}{3} p^3 (1-p)^2 + \binom{5}{4} p^4 (1-p) + \binom{5}{5} p^5.$$

The probability that that block does not give rise to a decoding error is $1 - p'$, so the probability that there is

no decoding error is $(1 - p')^k$ and the probability that there is a decoding error is $1 - (1 - p')^k$. If we fix $p > 0$ and let $k \rightarrow \infty$, then $(1 - p')^k$ tends to 0 (exponentially quickly), so the probability of decoding error tends to 1. Thus, this encoding/decoding pair is highly unreliable, and its rate is not too good either. The only redeeming feature is that it is very easy indeed to compute. (Its computational efficiency is easily seen to be bounded by a number of operations that is linear in k .)

One way to salvage the repetition code is to repeat every symbol $c \log k$ times. For a largish constant c , the probability of a decoding error goes to 0, but now the rate of the code goes to 0 as well. Prior to the work of Shannon it may have even been believed that a trade-off of this kind was inevitable: every encoding/decoding scheme would either achieve a vanishingly small rate or make mistakes with probability tending to 1. As we will see later in the article, it is in fact possible to define encoding schemes that achieve all three of our goals: they operate at a positive rate, they can correct errors that occur a positive proportion of the time (in either the probabilistic or the worst-case model), and they use efficient encoding and decoding algorithms. Most of the insight for this remarkable result goes back to a seminal paper by Shannon (1948). In that paper he gave the first examples of encoding and decoding functions that satisfied the first two goals, though they were not computationally efficient.

Shannon's encoding and decoding functions were therefore not practical, but we can now see, with the benefit of hindsight, that ignoring the goal of efficient computability in order to gain some theoretical insight into the channels was extraordinarily fruitful. A general rule of thumb seems to operate: that the performance of the very best encoding and decoding functions can be matched arbitrarily closely by encoding and decoding functions that are also computationally efficient. This justifies considering the goal of efficiency separately from the other two goals.

3 The Existence of Good Encoding and Decoding Functions

In this section we will describe results that demonstrate the existence of encoding and decoding functions that have an extremely good rate and reliability. In order to describe these results, first proved by Shannon, it will be useful to consider two related notions introduced by Hamming in work that was essentially concurrent with that of Shannon.

In order to understand these notions, let us start by describing what makes one encoding function E better or worse than another. The task of the *decoding* function is to work out, when it receives a string y , what the original message m was. Notice that this is equivalent to working out what the encoded message $E(m)$ was, since no two messages are encoded in the same way. The possible encoded messages are called *codewords*: that is, a codeword is a string of length n that arises as $E(m)$ for some message $m \in \Sigma^k$.

What we are worried about is the possibility of confusing two codewords after errors have been introduced, and this depends only on the set of codewords, and not on which codeword corresponds to which original message. Therefore, we adopt what at first seems a strange definition: an *error-correcting code* is any set of strings of length n in the alphabet Σ (that is, any subset of Σ^n). The strings in an error-correcting code are still called codewords. This definition completely ignores the actual process of encoding of a message, but that is so that we can focus on the rate and the decoding error while ignoring computational efficiency. If we are given an encoding function E , then the corresponding error-correcting code is simply the set of all the codewords of E . Mathematically, this is just the image of the function E .

What makes an error-correcting code good or bad? To answer this question, let us consider what happens if the alphabet is $\{0, 1\}$ and the code contains two strings $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$ that differ in precisely d places. If errors are introduced with probability p , then the probability that x is converted into y is $p^d(1 - p)^{n-d}$. Assuming that $p < \frac{1}{2}$, this probability gets smaller as d increases, so the smaller d is, the more likely the strings x and y are to be confused. It seems preferable, therefore, that there should not be too many pairs of strings in the code that differ in just a few places. A similar argument applies to larger alphabets as well.

The above thoughts lead to a definition that is very natural in this context. Given an alphabet Σ and two strings $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$ belonging to Σ^n , the *Hamming distance* between x and y is defined to be the number of coordinates i for which $x_i \neq y_i$. For example, let $\Sigma = \{a, b, c, d\}$ and let $n = 6$. The strings *abccad* and *abdcab* differ in the third and sixth places and are identical otherwise, so their Hamming distance is 2. Our goal is to find an encoding function E such that the associated code maximizes the typical Hamming distance between pairs of codewords.

Shannon's solution to this is an extremely simple application of the PROBABILISTIC METHOD [IV.19 §3]: he picks the encoding function at random. That is, for every message m , the encoding $E(m)$ is chosen entirely randomly from the set Σ^n , with all choices equally likely. Furthermore, for every message m , this choice is independent of the encoding of every other message m' . It is a good exercise in basic probability to see that such a choice almost always leads to a code where the distances between codewords are on average large. In fact, even the minimum distance between codewords is almost always large. However, we will not show this. Instead, we will argue that with high probability this random choice leads to a "nearly optimal" encoding function, from the point of view of rate and reliability.

First, let us consider what the decoding function ought to be. In the absence of computational requirements, it is not hard to say what the "optimal" decoding algorithm is. If you receive a sequence z , then you should choose the message m that is most likely to have resulted in this sequence. For the model Σ -SC(p) with $p < 1 - 1/|\Sigma|$, it is easily verified that this will be the message m for which the encoding $E(m)$ is nearest to z , as measured by Hamming distance. (If the minimum distance is attained by both $E(m)$ and $E(m')$, then one can make an arbitrary choice between them.) The condition on p is important here. It ensures that when the sequence $E(m)$ passes through the channel, the most likely output corresponding to any given term, out of the $|\Sigma|$ different possibilities, is the same as the input. Without this condition, there would be no reason to expect z to be close to $E(m)$. We shall argue that there is a number C , depending only on the error probability p and the size of the alphabet, such that for a random encoding function with rate smaller than C , this decoding function recovers the original message with a high probability. As an aside, Shannon also showed that for the same constant C , any attempt to communicate at rates greater than C would lead to errors with probability exponentially close to 1. Because of this result, the constant C is known as the *Shannon capacity* of the channel.

Once again, for simplicity we shall consider just the case of the binary alphabet $\{0, 1\}$. In this case we are choosing a random function E from $\{0, 1\}^k$ to $\{0, 1\}^n$, and we would like to show that, under suitable circumstances, the resulting code will almost certainly be very reliable. In order to do this, we shall focus on a single message m , and rely on two basic ideas.

The first idea is a precise form of THE LAW OF LARGE NUMBERS [III.73 §4]. If the error probability is p , then the expected number of errors introduced into a codeword $E(m)$ is pn , so, if n is large, then we expect that the actual number of errors will almost certainly be very close to this, just as, if you toss a fair coin ten thousand times, you will be surprised if the number of heads is not close to five thousand. The result that expresses this formally is as follows.

Claim. There exists a constant $c > 0$ such that the probability that the number of errors exceeds $(p + \epsilon)n$ is at most $2^{-c\epsilon^2 n}$.

The same can be said of the probability that the number of errors is less than $(p - \epsilon)n$, but we shall not use this result.

When n is large, $2^{-c\epsilon^2 n}$ is extremely small, so the number of errors is almost certainly at most $(p + \epsilon)n$. The number of errors equals the Hamming distance from y , the output of the channel, to $E(m)$, the codeword that was transmitted. Therefore, the decoding function that chooses the codeword with smallest Hamming distance from y will almost certainly choose $E(m)$, provided that there is no message m' such that $E(m')$ is closer to y than $(p + \epsilon)n$.

The second idea, which allows us to say that this will almost certainly be the case, is that "Hamming balls are small." Let z be a sequence in $\{0, 1\}^n$. Then the *Hamming ball of radius r about z* is the set of all sequences w with Hamming distance at most r from z . How big is this set? Well, in order to specify a sequence w with Hamming distance exactly d from z , it is enough to specify the set of d places where w and z differ. There are $\binom{n}{d}$ ways of choosing this set, so the number of sequences at a distance of at most r is

$$\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{r}.$$

If $r = \alpha n$ and $\alpha < \frac{1}{2}$, then this number is at most a constant times $\binom{n}{r}$, because each term is at least

$$\frac{n-r}{r} = \frac{1-\alpha}{\alpha}$$

times the one before. But

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

If we now use STIRLING'S FORMULA [III.31] or the looser approximation $n! = (n/e)^n$, then we find that this is about $(1/\alpha(1-\alpha))^n$, which is $2^{H(\alpha)n}$, where

$$H(\alpha) = -\alpha \log_2 \alpha - (1-\alpha) \log_2 (1-\alpha).$$

Terri: Tim thinks this is OK and could not possibly cause confusion.

(Note that $H(\alpha)$ is positive, because α and $1 - \alpha$ are less than 1 and therefore have negative logarithms.) The function H is called the *entropy function*. It is continuous and strictly increasing on the interval $[0, \frac{1}{2}]$ with $H(0) = 0$ and $H(\frac{1}{2}) = 1$. So, if $\alpha < \frac{1}{2}$, then $H(\alpha) < 1$, and therefore $2^{H(\alpha)n}$ is exponentially smaller than 2^n : this is what is meant by saying that the Hamming ball of radius αn is small.

Let us set α to be $p + \epsilon < \frac{1}{2}$. Then the probability that a single randomly chosen sequence $E(m')$ lies in the Hamming ball of radius $(p + \epsilon)n$ about y is at most $2^{H(p+2\epsilon)n} 2^{-n}$. (The 2ϵ is to compensate for slight inaccuracies in the above estimate for the size of the ball.) Since there are $2^k - 1$ possibilities for m' , the probability that one can be found for which $E(m')$ lies in the ball is at most $2^k 2^{H(p+2\epsilon)n} 2^{-n}$. Therefore, if $k \leq n(1 - H(p + 2\epsilon) - \epsilon)$, this probability is at most $2^{-\epsilon n}$, which is exponentially small.

Because we can choose ϵ to be as small as we like, we can make k/n as close as we like to $1 - H(p)$ while still maintaining an exponentially small probability of decoding error. It turns out that the quantity $1 - H(p)$ is the constant C discussed earlier: the Shannon capacity of the binary symmetric channel. Thus, the capacity of the binary symmetric channel is always positive if $p < \frac{1}{2}$.

Shannon's theorem and proof are significantly more general than the above example demonstrates. For a wide variety of channels, and for a wide variety of models of (probabilistic) error, his theory pins down the capacity of the channel and shows that reliable communication is possible if and only if the rate of the channel is less than its capacity. Shannon's proof is a remarkable example of the use of the probabilistic method in the practice of engineering. Note, however, that the encoding and decoding algorithms are quite impractical. The proof gives no clue about how to find an encoding function, though of course one can consider every encoding function $E : \{0, 1\}^k \rightarrow \{0, 1\}^n$ to check if it is good. However, even if such a function is found, it may have no succinct description, in which case the encoder and decoder have to store this encoding function as an exponentially long table in their memory. Finally, the decoding algorithm seems to involve a brute-force search for the nearest codeword, a problem which seems to be the most serious obstacle to obtaining a computationally efficient version of Shannon's theorem that can be used in practice. What the theorem definitely *does* give us is a significant insight into the limitations and potential utility of the

communication channel. With this in mind, we can set ourselves the right targets to strive for when we come to devise more practical encoding and decoding procedures. In the next section we will show that it is possible to achieve a fixed rate that is bounded away from zero, to tolerate a constant fraction of errors, and to do both of these with efficient algorithms.

4 Efficient Encoding and Decoding

We now turn to the task of designing encoding and decoding functions that can be calculated efficiently. Currently, there are at least two very different approaches to building such functions. We describe here an approach based on algebra over finite fields. The alternative approach is based on the construction of EXPANDING GRAPHS [III.24], but we will not describe that here.

4.1 Codes for Large Alphabets Using Algebra

In this section we describe a simple way to get an encoding function $E : \Sigma^k \rightarrow \Sigma^n$, where Σ is a finite FIELD [I.3 §2.2] with at least n elements. (Recall that there are finite fields with q elements whenever q is of the form p^t for a prime p and a positive integer t .) These codes were introduced by Reed and Solomon (1960) and have since been called the *Reed-Solomon codes*.

A Reed-Solomon code is specified by a sequence of n distinct field elements $\alpha_1, \dots, \alpha_n \in \Sigma$. Given a message $m = (m_0, m_1, \dots, m_{k-1}) \in \Sigma^k$, we associate with the message the polynomial $M(x) = m_0 + m_1x + \dots + m_{k-1}x^{k-1}$. The encoding of m is simply the sequence $E(m) = (M(\alpha_1), M(\alpha_2), \dots, M(\alpha_n))$. In other words, to encode a sequence m , you treat the terms of the sequence as the k coefficients of a polynomial of degree $k - 1$ and write out the values that this polynomial takes at $\alpha_1, \dots, \alpha_n$.

Before describing the error-correcting capability of this code, let us note that it is very succinctly represented: all that is needed to specify it is a description of the field Σ and the sequence of n elements $\alpha_1, \dots, \alpha_n$. It is easy to show that the number of additions and multiplications needed to compute $M(\alpha)$ is at most Ck for some constant C . (For example, to work out $3\alpha^3 - \alpha^2 + 5\alpha + 4$, you start with 3, multiply by α , subtract 1, multiply by α , add 5, multiply by α , and add 4.) Therefore, the number of field operations needed to compute the entire encoding is bounded above by Cnk , for some (different) constant C . (In fact, more sophisticated and efficient algorithms are known

for the encoding problem that take at most $Cn(\log n)^2$ steps.)

Now let us consider the error-correcting properties of the code. We start by showing that the encodings of any two messages m_1 and m_2 have a Hamming distance of at least $n - (k - 1)$. To see this, let $M_1(x)$ and $M_2(x)$ be the polynomials associated with m_1 and m_2 . Now the difference $p(x) = M_1(x) - M_2(x)$ has degree at most $k - 1$, and it is not the zero polynomial (since M_1 and M_2 are distinct), and therefore it has at most $k - 1$ roots. This tells us that there are at most $k - 1$ values of α for which $M_1(\alpha) = M_2(\alpha)$. It follows that the Hamming distance between the sequences

$$E(m_1) = (M_1(\alpha_1), M_1(\alpha_2), \dots, M_1(\alpha_n))$$

and

$$E(m_2) = (M_2(\alpha_1), M_2(\alpha_2), \dots, M_2(\alpha_n))$$

is at least $n - k + 1$.

It follows that if z is any sequence, then its Hamming distance from at least one of $E(m_1)$ and $E(m_2)$ is greater than $\frac{1}{2}(n - k)$ (since otherwise the distance between $E(m_1)$ and $E(m_2)$ would have to be at most $n - k$). Therefore, if the number of errors that occur during transmission is at most $\frac{1}{2}(n - k)$, then the original message m is uniquely determined by the received sequence z . What is much less obvious is that there is an efficient algorithm for working out what m was, but, remarkably, it is possible to compute m with a polynomial-time algorithm (in n), which we shall now describe.

What must the decoding algorithm do? It is given the numbers $\alpha_1, \dots, \alpha_n$ and the received sequence z_1, \dots, z_n and is required to find a polynomial M of degree $k - 1$ or less such that $M(\alpha_i) = z_i$ for all but at most $\frac{1}{2}(n - k)$ values of i . If such a polynomial exists, then it is unique, as we have just seen, and its coefficients will give the original message m (if the number of errors is at most $\frac{1}{2}(n - k)$).

If there were no errors, then our task would be much easier: one can determine the coefficients of a polynomial of degree $k - 1$ from k of its values by solving k simultaneous equations. However, if some of the values we use are incorrect, then we will end up with a completely different polynomial, so this method is not easy to use for the problem we actually face.

To overcome this difficulty, let us imagine that M exists and that the errors introduced into the sequence $M(\alpha_1), \dots, M(\alpha_n)$ occur at i_1, \dots, i_s , where $s \leq \frac{1}{2}(n - k)$. Then the polynomial $B(x) = (x - \alpha_{i_1}) \cdots$

$(x - \alpha_{i_s})$ has degree at most $\frac{1}{2}(n - k)$ and is zero if and only if x is equal to α_{i_j} for some j . Let us set $A(x)$ to equal $M(x)B(x)$. Then $A(x)$ is a polynomial of degree at most $k - 1 + \frac{1}{2}(n - k) = \frac{1}{2}(n + k - 2)$, and for every i we have $A(\alpha_i) = z_i B(\alpha_i)$. (If there is no error at i , then this is obvious, since $z_i = M(\alpha_i)$, and if there is an error at i , then both sides are 0.)

Conversely, suppose that we manage to find polynomials $A(x)$, of degree at most $\frac{1}{2}(n + k - 2)$, and $B(x)$, of degree at most $k - 1$, such that $A(\alpha_i) = z_i B(\alpha_i)$ for every i . Then $R(x) = A(x) - M(x)B(x)$ is a polynomial of degree at most $\frac{1}{2}(n + k - 2)$, and $R(\alpha_i) = 0$ whenever $M(\alpha_i) = z_i$. Since there are at most $\frac{1}{2}(n - k)$ errors, this happens for at least $n - \frac{1}{2}(n - k) = \frac{1}{2}(n + k)$ values of i . Therefore, the number of roots of R is bigger than its degree, from which it follows that R is identically zero, so that $A(x) = M(x)B(x)$ for every x . From this we can determine M : given k values of x for which $A(x)$ and $B(x)$ are nonzero, one can determine k values of $M(x) = A(x)/B(x)$, and hence determine M .

It remains to show that we can indeed (efficiently) find polynomials $A(x)$ and $B(x)$ with the required properties. The n constraints $A(\alpha_i) = z_i B(\alpha_i)$ turn into n linear constraints on the unknown coefficients of A and B . Since B has $\frac{1}{2}(n - k) + 1$ coefficients and A has $\frac{1}{2}(n + k)$ coefficients, the total number of unknowns is $n + 1$. Since the system of equations is homogeneous (that is, we obtain a solution if we take all unknowns to be zero) and the number of unknowns is greater than the number of constraints, there must be a nontrivial solution: that is, a solution where $A(x)$ and $B(x)$ are not both the zero polynomial. Moreover, we can find such a solution by Gaussian elimination, which takes at most Cn^3 steps.

To summarize: we construct a code by exploiting the fact that two distinct low-degree polynomials cannot be equal for too many values. We then exploit the rigid algebraic structure of low-degree polynomials for the purposes of decoding. The main tool that allows us to do this is linear algebra and in particular the solving of systems of simultaneous equations.

4.2 Reducing the Size of the Alphabet Using Good Codes

The ideas described in the previous section show us how to build codes with efficient encoding and decoding algorithms, but they use relatively large alphabets. In this section we shall exploit these results to build binary codes.

To begin with, let us consider a very obvious method of converting codes over large alphabets into codes over the binary alphabet $\{0, 1\}$. For simplicity, assume that we have a Reed-Solomon code over an alphabet Σ of size 2^l for some integer l . Then we can associate the elements of Σ with binary strings of length l . In such a case, we can regard the Reed-Solomon encoding function, which maps Σ^k to Σ^n , as a function from $\{0, 1\}^{lk}$ to $\{0, 1\}^{ln}$. (For instance, an element of Σ^k is a sequence of k objects, each of which is a binary sequence of length l . Putting them together produces a single binary sequence of length kl .) Since the encodings of two distinct messages differ for at least $n - k + 1$ elements of Σ , they must also differ on at least $n - k + 1$ bits.

This gives a fairly reasonable code over the binary alphabet. However, $n - k + 1$ is not as large as a fixed fraction of ln : the ratio $(n - k + 1)/ln$ is less than $1/l$, and since we need 2^l , the size of Σ , to be at least n , we find that this fraction is at most $1/\log_2 n$, which tends to zero as n tends to infinity. However, this can be fixed in a simple way, as we shall see.

The problem with the simple binary approach is that two different elements of Σ may be represented by binary sequences that differ in just one bit. However, the Hamming distance between two binary sequences of length l is usually much larger: it is more like cl for some positive constant c . Suppose that we could represent the elements of Σ as binary sequences of some length L in such a way that the Hamming distances between any two of the sequences used was at least cL . This would allow us to improve our argument above: if the encodings of two messages were different for at least $n - k + 1$ elements of Σ , then they would have to differ on at least $cL(n - k + 1)$ bits rather than just $n - k + 1$, and this is a positive fraction of Ln .

What we are asking for is an encoding of the binary sequences of length l as sequences of length L in such a way that no two codewords are closer than cL to each other. But we know, from the previous section, that such an encoding exists, provided that L and c satisfy appropriate conditions: for instance, it is possible to find an encoding function that works with $L \leq 10l$ and $c \geq \frac{1}{10}$.

So how do we use this? We start with a binary sequence m of length lk . As above, we associate with this a sequence of length k in the alphabet Σ . We then encode this sequence using the Reed-Solomon code, obtaining a sequence of length n in the alphabet Σ . Next, we convert each term of this sequence into a binary sequence of length l . And, finally, we encode

each of these n binary sequences as a sequence of length L using a good encoding function, obtaining as a result a binary sequence of length Ln . We then pass this sequence through the channel, where errors may be introduced. The receiver then breaks the received sequence up into n blocks of length L , decodes each block to work out what binary sequence of length l gave rise to it, and interprets that binary sequence as an element of Σ . This results in a sequence of n elements of Σ . It then uses the Reed-Solomon decoding algorithm to decode this sequence, producing a sequence of k elements of Σ . Finally, this can be converted into a binary sequence of length lk .

We have said nothing about the efficiency of the encoding and decoding procedures that convert binary sequences of length l into ones of length L and back again, stating merely that they exist. Since efficiency is supposed to be our priority, this may seem rather strange: do we not now face exactly the same problem that we were trying to solve in the first place? Luckily we do not, because although these encoding and decoding procedures may take exponentially long, they take exponentially long as a function of L , and L is much much smaller than n . Indeed, L is proportional to $\log n$, from which it follows that 2^L is bounded above by a polynomial function of n . This is a useful principle: one can afford procedures of exponential complexity provided that one only ever applies them to very short strings.

Thus even though we have not managed to specify the code explicitly, we have demonstrated that there is an encoding and decoding algorithm that runs in polynomial time and that corrects a constant fraction of errors. To complete this section, let us address the question of the probability of decoding error, which we have not yet discussed. The technique described above, of composing encoding functions (and decoding functions), can also be used to improve the above code so that the encoding and decoding still take place in polynomial time, but now the decoding error probability is exponentially small on the binary symmetric channel with parameter p , and the rate is arbitrarily close to the Shannon capacity, which is the theoretical maximum. (The idea is to compose a Reed-Solomon code that has rate close to 1 with a random inner code, and then to show that with random errors most of the inner decoding steps decode correctly. One then uses the outer decoding step to convert the “mostly correct decoding” to a “fully correct decoding.”)

5 Impact on Communication and Storage

The mathematical theory of error-correcting codes has made a deep impact on the technologies for storage and communication of information, and we elaborate a little on this below.

Storage of information on digital media is probably the biggest success story for error-correcting codes. Most known forms of storage media, and in particular standards for audio and data CDs and DVDs, prescribe error-correcting codes based on Reed-Solomon codes. Specifically, they are based on a code that maps \mathbb{F}_{256}^{223} to \mathbb{F}_{256}^{255} , where \mathbb{F}_{256} is the finite field with 256 elements. In audio CDs, codes are used to protect from minor scratches, though more serious scratches do lead to audible errors. In data CDs the error correction is stronger (with more redundancies), so that even serious scratches do not lead to loss of data. In all cases (CDs and DVDs) the readers for these devices use fast algorithms for decoding when reading the information on the media. Typically, these algorithms are based on the idea of the previous section, but are much faster implementations (in particular, an algorithm due to E. Berlekamp is widely used). Indeed, several CD readers owe their faster reading speed to faster decoding algorithms. Similarly, the increased storage capacity of DVDs (compared with CDs) is attributed in part to better error-correcting codes. Indeed, error-correction technology played a crucial role in establishing the dominance of audio CDs, which store music digitally, over the traditional, and now almost extinct, gramophone records, which store music in continuous forms. Thus, mathematical advances in coding theory have played an influential role in this technology.

Similarly, error-correcting codes have had a profound effect on communication. Since the late 1960s, error-correcting codes (and decoding) have been used for communication from satellites to their base stations on Earth. Of late, error-correcting codes are also being used in cellular phone communications and modems. Again, the most commonly used code at the time of the writing of this article is the Reed-Solomon code, though this situation has been changing rapidly since the discovery of a new class of codes called “turbo codes.” This new family of codes seems to offer significant resilience to random errors (more so than that offered by methods based on Reed-Solomon codes) and uses a simple and quick algorithm, even when the codes used have small block length. These codes and the corresponding decoding algorithm have led to a resurgence of inter-

est in codes constructed with the help of insights from GRAPH THEORY [III.34]. Many of the good properties of turbo codes have been observed only empirically: that is, the codes seem to work very well in practice but it has not yet been proved rigorously that they do. Nevertheless, the observations have been so compelling that new standards for communication are starting to prescribe these codes.

Finally, it must be stressed that while many of the codes used are based on ones that are studied in the mathematical literature, this should not be taken to mean that they can be deployed immediately without further design. For example, the *Mariner* spacecraft used not a Reed-Muller code but a variant of it designed to allow for synchronization between blocks. Similarly, the Reed-Solomon codes used in storage devices are carefully spread out over the disc, so as to allow the physical device to resemble more closely the model of a code over a large alphabet. Note that errors due to a scratch on the disc surface tend to ruin a large collection of bits in a small localized part of the disc. If all the data from a block were sitting in such a neighborhood, the entire block would be lost. So each block of 255 bytes of information is spread out all over the disc. On the other hand, the bytes themselves, which are elements of \mathbb{F}_{256} , are written as eight bits in close proximity. So a scratch corrupting one bit out of these eight is also likely to corrupt others in the neighborhood. However, this is all right from the perspective of the model that views the entire collection of eight bits as a single element. In general, working out the right way to apply the theory of error correction to a given scenario is a major challenge, and many success stories would not have been success stories had it not been for some careful design choices.

Mathematics and engineering continue to feed each other in this arena. Mathematical successes, such as new algorithms for decoding Reed-Solomon codes, raise the challenge of how to adapt technology to exploit new algorithms. Engineering successes, such as the discovery of turbo codes that perform extremely well, challenge mathematicians to come up with a formal model and analysis that can explain this success. And if such a model and analysis emerges, it is likely to lead to the discovery of new codes that might surpass the performance of turbo codes and lead to a new set of standards!

6 Bibliographic Notes

The theory of reliable communication and storage of information owes much to the seminal works of Shannon (1948) and Hamming (1950), which formed the basis for much of this article. The Reed–Solomon codes of section 4.1 are from Reed and Solomon (1960). Their decoding algorithm originates in the work of Peterson (1960), though the algorithm given here is significantly simplified. The technique of composing codes is due to Forney (1966).

Over the years, coding theory has amassed a wide variety of results. Some of these give better constructions of codes with faster algorithms. Others provide theoretical upper limits on how well codes can perform. The theory uses an enormous variety of mathematical tools, many of them more advanced than the ones described in this article. Most notable among them are algebraic geometry and graph theory, which are used to construct very good codes, and the theory of orthogonal polynomials, which is used to prove limits on parameters of codes, such as their rate and reliability. Most of the highlights of this vast literature are covered in Pless and Huffman (1998).

Further Reading

- Hamming, R. W. 1950. Error detecting and error correcting codes. *Bell System Technical Journal* 29:147–60.
- Forney Jr., G. D. 1966. *Concatenated Codes*. Cambridge, MA: MIT Press.
- Peterson, W. W. 1960. Encoding and error-correction procedures for Bose–Chaudhuri codes. *IEEE Transactions on Information Theory* 6:459–70.
- Pless, V. S., and W. C. Huffman, eds. 1998. *Handbook of Coding Theory*, two volumes. Amsterdam: North-Holland.
- Reed, I. S., and G. Solomon. 1960. Polynomial codes over certain finite fields. *SIAM Journal of Applied Mathematics* 8:300–4.
- Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27:379–423, 623–56.

VII.7 Mathematics and Cryptography

Clifford Cocks

1 Introduction and History

Cryptography is the science of hiding the meaning or content of communications. The aim is that an adversary who sees a message only in its enciphered state cannot make sense of or derive useful information from

what is seen. On the other hand, the intended recipient must be able to decipher the true meaning. For most of history cryptography has been an art practiced seriously only by a few—such as governments for military and diplomatic communications—for whom the consequences of unauthorized disclosure of information are damaging enough to justify the expense and inconvenience of enciphering messages. Recently this has changed: one of the results of the information revolution has been the need for instant and secure communication for all on demand. Fortunately, mathematics has come to the rescue and provided theoretical and algorithmic developments to meet this need. It has also provided entirely new possibilities, such as “digital signatures” (which will be discussed later).

One of the oldest and most basic methods of cryptography is *simple substitution*. Suppose that a message to be enciphered consists of a piece of English text. Before it is sent, the sender and recipient agree on a permutation of the twenty-six letters of the alphabet, which they keep private. An enciphered message might then look something like

ZPLKKWL MFUPP UFL XA EUXMFLP

For very short messages this method is reasonably secure—it is just possible to work out the meaning of the above example by matching letter patterns to those commonly seen in English, but it is quite challenging! However, for longer messages, simply counting the frequencies of each letter and comparing those counts with the frequencies of letters in natural language will almost always reveal the hidden permutation sufficiently to allow the meaning to be easily recovered.

A major leap forward in cryptography came with the advent of mechanical encryption devices in the twentieth century, of which the German Enigma used during World War II is perhaps the most famous example. An account of the fascinating Enigma story and the role of the code breakers of Bletchley Park appears in Simon Singh’s excellent book on cryptography (Singh 1999). It is interesting that the principle on which Enigma operates is a development of the simple substitution method. Each letter of the input message is enciphered exactly as a simple substitution, but with the additional rule that the permutation controlling the substitution changes after every letter. A complex electro-mechanical device controls the substitution process in a deterministic way. The recipient can decipher the message only if he or she can set up another device

Terri: Tim and the author (presumably) think it’s amusing not to include the solution. OK?

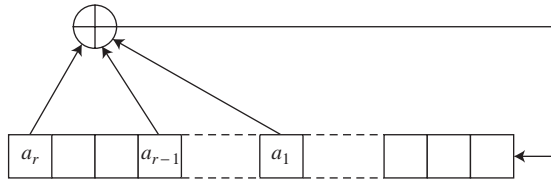


Figure 1 Linear feedback shift register.

in exactly the same way as the originator. The information needed to do this is called the *key*. Making sure that keys are known only by the right people is called *key management*. Until the advent of public-key cryptography (to be discussed later), key management was a major inconvenience and expense for anyone wanting to secure their communications.

Terri: Tim would rather keep 'anyone' here. OK?

2 Stream Ciphers and Linear Feedback Shift Registers

Since the advent of computers, information has tended to be transmitted as *binary data*: that is, as a stream of 0s and 1s. For such data there is a rather different method of encipherment based on a device called the *linear feedback shift register*, or LFSR (see figure 1). The first step is to generate a random-looking sequence of 0s and 1s in a deterministic way, and this is done by means of a recurrence formula, of which a simple example is

$$x_t = x_{t-3} + x_{t-4}.$$

Here, addition is mod 2, so x_t will be 1 if an odd number of the terms x_{t-3} , x_{t-4} is 1, and it will be 0 otherwise. We must also specify the first four values of the sequence, so let us begin with 1000. The sequence then continues as follows:

100110101111000100110101111...

More generally, one specifies some positive integers a_1, a_2, \dots, a_r , called *feedback positions*—the numbers 3 and 4 in the above example—and defines a sequence by means of the recurrence formula

$$x_t = x_{t-a_1} + x_{t-a_2} + \dots + x_{t-a_r},$$

where again the addition is mod 2.

A sequence produced in this way usually looks fairly random, but because there are only finitely many binary sequences of length a_r it must eventually repeat. Notice that, in our example, the sequence is periodic with period 15, which is actually the longest possible period, since there are sixteen binary sequences of length 4, and after a moment's thought one sees

that the sequence 0000 cannot occur (or else the whole sequence up to then would have had to consist entirely of zeros).

In general, the length of the sequence depends on properties of the polynomial

$$P(x) = 1 + x^{a_1} + x^{a_2} + \dots + x^{a_r}$$

over the FIELD [I.3 §2.2] \mathbb{F}_2 of two elements. As we have just seen in the case $a_r = 4$, the maximum possible sequence length is $2^{a_r} - 1$, and for this length to be achieved the polynomial $P(x)$ must be *irreducible* over \mathbb{F}_2 : that is, it must not factorize into smaller polynomials. For example, the polynomial $1 + x^4 + x^5$ is not irreducible, because $(1 + x + x^3)(1 + x + x^2)$ expands out to

$$1 + x + x + x^2 + x^2 + x^3 + x^3 + x^4 + x^5,$$

which equals $1 + x^4 + x^5$ since $1 + 1 = 0$ in the field \mathbb{F}_2 .

Irreducibility is a necessary condition for the sequence to have the maximum length, but it does not guarantee it. For that we need a second condition: that the polynomial is *primitive*. To see what this means, let us take the polynomial $x^3 + x + 1$ and calculate the remainder when, for the first few positive integers m , we divide x^m by $x^3 + x + 1$ (with all coefficients in \mathbb{F}_2). When m goes from 1 to 7 we obtain the polynomials x , x^2 , $x + 1$, $x^2 + x$, $x^2 + x + 1$, $x^2 + 1$, 1. For instance,

$$x^6 = (x^3 + x + 1)(x^3 + x + 1) + x^2 + 1,$$

so the remainder on dividing x^6 by $x^3 + x + 1$ is $x^2 + 1$.

Now the first time that we obtained the polynomial 1 was when $m = 7$, and $7 = 2^3 - 1$. This shows that the polynomial $x^3 + x + 1$ is primitive. In general, a polynomial $p(x)$ of degree d is primitive if the first time you obtain a remainder of 1 when you divide x^m by $p(x)$ is when $m = 2^d - 1$.

There are computationally efficient tests for determining whether a polynomial is irreducible and whether it is primitive. The advantage of using a primitive polynomial as the basis of an LFSR is that, in the sequence it generates, no subsequence of length a_r is repeated until all nonzero sequences of length a_r have appeared exactly once.

How is all this applied in cryptography? A simple idea would be to take the stream of bits generated by an LFSR and add it term by term to the message one is enciphering. For instance, if the LFSR generated a sequence that began 1001101 and the message was 0000111, then the encrypted message would begin 1001010. To decipher such a message, one could simply repeat the process: adding the two sequences 1001101 and

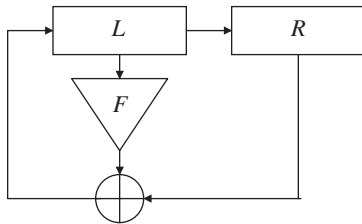


Figure 2 Feistel round structure.

1001010 gives the original message 0000111. For this to work, the recipient would need to know the details of the LFSR in order to be able to generate the same sequence 1001101, so one might consider using the feedback positions (in this case 3 and 4) as the secret key.

In practice, the above procedure is not good enough because there is an efficient algorithm, due to Berlekamp and Massey (1969), that can recover the feedback rule from the stream of bits it generates. It is better to use some predetermined nonlinear function of the successive sequences of a_r bits in order to scramble further the sequence of bits produced by the LFSR. Even then, such procedures are simple enough that, with careful design, they can be applied to large amounts of data very quickly.

3 Block Ciphers and the Computer Age

3.1 Data Encryption Standard

When computers started to be used, an entirely different method of cryptography became practical: the block cipher. The first example of this was DES: the *Data Encryption Standard* (first published in 1977). DES was adopted as a standard in 1976 by the U.S. National Bureau of Standards (now the National Institute of Standards and Technology). This enciphers a block of 64 bits at a time, with a key of length 56 bits. It has a particular structure, referred to as a *Feistel cipher* (see figure 2).

This structure is as follows. Given a block of 64 bits, you first divide it into two parts of 32 bits each, and call them L and R . Next, you take a subset of the 56 bits of the key, according to some predetermined rule, and use this subset to define a nonlinear function F , again according to some predetermined rule, which takes 32-bit sequences to 32-bit sequences. You then replace the pair $[L, R]$ by the pair $[R \oplus F(L), L]$. (Here $R \oplus F(L)$ denotes the result of taking the mod-2 sum of R and $F(L)$ one bit at a time.)

Having done that, you repeat the process a number of times, choosing a different nonlinear function F each time (but always deriving it in a predetermined way from the 56-bit key). A complete encryption by DES consists of 16 such rounds, together with some permutation of the bits of the input and output.

One reason for using the Feistel structure is that as long as one knows the 56-bit key it is quite easy to reverse the encryption process. Given a round that performs the transformation

$$[L, R] \rightarrow [R \oplus F(L), L],$$

one can invert it by means of the transformation

$$[L, R] \rightarrow [R, L \oplus F(R)].$$

This has the great advantage that it does not require us to invert F , so even if F is quite complicated the procedure can be easy to carry out.

A number of what are called “modes of use” of DES have been developed. Simply using the algorithm to encrypt each 64-bit block of data in turn is called ECB (for *electronic codebook*) mode. A disadvantage of this mode is that if there is an exact 64-bit repeat in the data then this results in an exact 64-bit repeat in the cipher.

Another mode is CBC, or *cipher block chaining*, mode. Here, each block of data is added mod 2 to the previous block before being encrypted as above. In OFB, or *output feedback*, mode the block of data is added to the DES encipherment of the previous block. It is an easy exercise to see how to decipher in CBC and OFB modes, and in practice these are the two most common modes of use of DES.

3.2 Advanced Encryption Standard

The U.S. National Institute of Standards and Technology recently held a competition for a replacement for DES, to be called the *Advanced Encryption Standard*, or AES. This was to be a 128-wide block cipher with a variety of possible key lengths. Many competing designs were submitted and subjected to public scrutiny, and the winning entry was called *Rijndael*, after the designers Joan Daemen and Vincent Rijmen.

The design is remarkable and elegant and makes use of interesting mathematical structures (Daeman and Rijmen 2002). The 128 bits in each block are thought of as 16 bytes (a byte consists of eight bits), arranged in a 4×4 square. Each byte is then thought of as an element of \mathbb{F}_{256} , the field of order 256. Encryption consists of ten or more rounds (the exact number depending upon

the key length); and each round mixes the data and the key.

A round consists of a series of steps, typically as follows. First, each byte, regarded as an element of the finite field \mathbb{F}_{256} , is replaced by its inverse in the field, except that 0 is left unchanged. Each byte is then regarded as an element of the vector space of dimension 8 over the field \mathbb{F}_2 and an invertible linear transformation is applied. Each row of the 4×4 square is then rotated, by a different number of bytes for each row. Next, the values of each column of the square are taken to be the coefficients of a degree 3 polynomial over \mathbb{F}_{256} and this is multiplied by a fixed polynomial and reduced modulo $x^4 + 1$. Finally, the key for the round, which is derived linearly from the encryption key, is added modulo 2 to the 128 bits.

It can be seen that all of these steps are reversible, which makes decipherment straightforward. It is likely that AES will take over from DES as the most widely used block cipher.

4 One-Time Key

The various encryption methods described above rely on the computational difficulty of recovering some secret that protects the enciphered data. There is one classic encryption method that does not rely on this property. This is the “one-time key.” Imagine that the message to be enciphered is encoded as a sequence of bits (for example, the standard ASCII encoding that represents each character as eight bits). Suppose that ahead of time the sender and recipient have shared a sequence of random key bits r_1, \dots, r_n at least as long as the message. Suppose that the message bits are p_1, p_2, \dots, p_n .

The enciphered message is then x_1, x_2, \dots, x_n , where $x_i = p_i + r_i$. Here, as usual, addition is mod 2 addition in each bit. If the bits r_i are fully random, then knowing the sequence x_i gives no information whatsoever about the message sequence p_i . This system is called *one-time key*. It is very secure as long as the key is used only once. However, it is impractical to use this method except in very specialized situations because of the need for sender and recipient to share and keep safe possibly large quantities of key material.

5 Public-Key Cryptography

All of the examples of encryption methods that we have seen so far have had the following structure. Two communicators agree on an algorithm or method for

encryption. The choice of method (e.g., simple substitution, AES, or one-time key) can be made public without the security of the system being compromised. The two communicators also agree on a secret key in the form required by the chosen encryption method. This key needs to be kept secure and never revealed to any adversary. The communicators encipher and decipher messages using the algorithm and secret key.

This presents a major problem: how can the communicators securely share the secret key? It would be insecure to exchange this over the same system that they will later use to send enciphered messages. Until so-called public-key methods were discovered this issue limited the use of encryption to those organizations that could afford the physical security and separate communication channels necessary for distributing keys reliably.

The following remarkable, counterintuitive proposition forms the basis of public-key cryptography: *it is possible for two entities to communicate information in such a way that they start with no secret shared information; an adversary has access to all communications between them; at the end the entities have shared secret knowledge that the adversary is unable to determine.*

It is easy to see how useful such a capability could be. Consider, for example, someone making a purchase over the Internet. Having identified a product one wishes to buy the next step is to send personal information such as credit card details to the vendor. With public-key cryptography it is possible to do this in a secure manner straightaway.

How might public-key cryptography be possible? The structure of a solution was proposed by James Ellis in 1969,¹ with the first public description by Diffie and Hellman (1976). The critical idea is to use a function that is hard to invert unless you have an “inverse key” that helps you to do so.

More formally, a *one-way function* H is a mapping from a set X to itself, with the property that if you are told the value $y = H(x)$ for some $x \in X$, then it is computationally hard to determine x . The inverse key is a secret value, z , say, used in creating the function H , with the property that if you know z then it becomes computationally *easy* to recover x from $H(x)$.

We can use this to solve the problem of secure key exchange as follows. Let us suppose that Bob wishes to send some data securely to Alice. (Particularly useful would be a shared secret that they can use later as a

1. See “The possibility of secure non-secret digital encryption,” available at www.cesg.gov.uk/site/publications/media/possnse.pdf.

key for subsequent communications.) Alice begins by generating a one-way function H with an inverse key z . She then communicates the function H to Bob, but the inverse key remains her personal secret, which she reveals to no one—not even to Bob. Bob takes the data x that he wishes to send, computes $H(x)$, and returns the result of his computation to Alice. Because Alice has the inverse key z , she can reverse the function H and thereby recover x .

Now suppose that an adversary manages to read all the communications between Alice and Bob. Then the adversary will know the function H and the value $H(x)$. However, Alice has not communicated the inverse key z , so the adversary is faced with the computationally intractable problem of inverting H . Therefore, Bob has successfully transmitted the secret x to Alice without the adversary being able to work out what it is. (For a more precise idea of what computational intractability is and a further discussion of one-way functions, see COMPUTATIONAL COMPLEXITY [IV.20], especially section 7.)

It can be helpful to imagine the one-way function H as a padlock and the inverse key as the key that unlocks the padlock. Then if Alice wants to receive an enciphered message from Bob, she sends him her padlock, retaining the key. Bob locks (enciphers) the message into a box with the padlock, and returns it. Only Alice, who is in possession of the padlock key, can unlock (decipher) the message.

5.1 RSA

It is all very well to have such a framework, but it leaves open an obvious question: how can one produce a one-way function with an inverse key? The following method was published by Rivest, Shamir, and Adleman (1978). It relies on the fact that it is relatively easy to find large prime numbers and multiply them to produce a composite number, but it is much harder, if you are given that composite number, to determine its two prime factors.

To create a one-way function by their method, Alice first finds two large prime numbers P and Q . She then calculates the integer $N = PQ$ and sends it to Bob, together with another integer e called the *encryption exponent*. The values N and e are called the *public parameters* because it does not matter if an adversary knows what they are.

Bob then expresses the secret value x that he wishes to send to Alice as a number modulo N . Next, he computes $H(x)$, which is defined to be $x^e \bmod N$, that is,

the remainder when x^e is divided by N . Bob sends $H(x)$ to Alice.

Upon receipt of Bob's message, Alice needs to recover x from $x^e \bmod N$. This she can do by first calculating the number d that satisfies the equation

$$de \equiv 1 \pmod{(P-1)(Q-1)}.$$

To do this efficiently, Alice can use EUCLID'S ALGORITHM [III.22]. Notice, however, that this would not be possible if she did not know the values of P and Q . In fact, the ability to calculate the correct value of d can be shown to be equivalent to the ability to factorize N . The value of d is Alice's private key (or "inverse key" in the terminology above): it is the secret that can undo the encryption function H .

This is because $H(x)^d \bmod N$ can be shown to equal x . Indeed, the significance of the number $(P-1)(Q-1)$ is that it equals $\phi(N)$, the number of integers less than N and coprime to N . EULER'S THEOREM [III.60] states that $x^{\phi(N)} \equiv 1 \bmod N$ whenever x is coprime to N . Therefore, $x^{m\phi(N)} \equiv 1 \bmod N$ as well, so if de has the form $m\phi(N) + 1$, as we are assuming, then $H(x)^d \equiv x^{de} \equiv x \bmod N$. In other words, if you raise x to the power $e \bmod N$ and then raise that to the power $d \bmod N$ you get back to x . (An important point is that raising numbers to powers $\bmod N$ is computationally easy by the method of "repeated squaring." This is discussed in COMPUTATIONAL NUMBER THEORY [IV.3 §2].)

While it has not been proved that the only way for an adversary to defeat the RSA encryption system is to factorize N , no other general attack has been found. This has created interest in finding improved factorization methods. A number of new subexponential methods—elliptic curve factorization (Lenstra 1987), the multiple polynomial quadratic sieve (Silverman 1987), and the number field sieve (Lenstra and Lenstra 1993)—have been discovered in the years since the RSA algorithm was found. See COMPUTATIONAL NUMBER THEORY [IV.3 §3] for discussions of some of them.

5.1.1 Implementation Details

The security of the RSA system depends on the primes P and Q being large enough to make factorization hard. However, the larger they are, the slower the encryption process is. Thus, there is a trade-off between security and the speed of encryption. A typical choice that is often made is to use primes that are each of 512 bits.

For the deciphering method to work, the encryption exponent e must have no factors in common with either $(P - 1)$ or $(Q - 1)$. This assumption was needed when we applied Euler's theorem, and if it does not hold then the encryption function is not invertible. Values such as 17 or $2^{16} + 1$ are often used in practice, because making e small reduces the amount of computation needed to calculate the encrypted value $x^e \bmod N$. (These two values of e are also well-suited to calculation by repeated squaring.)

5.2 Diffie-Hellman

Another approach to generating a shared secret was published by Whitfield Diffie and Martin Hellman. In their protocol Alice and Bob jointly create a shared secret, which can then be used as the key for one of the conventional cryptographic systems such as AES. To do this, they agree on a large prime number P and a *primitive element* g modulo P , which means a number g such that $g^{P-1} \equiv 1 \bmod P$, but $g^m \not\equiv 1 \bmod P$ for any $m < P - 1$.

Alice then creates her own private key a , a number randomly chosen between 1 and $P - 1$, and calculates $g_a = g^a \bmod P$ and sends this to Bob.

Bob similarly creates his own private key b between 1 and $P - 1$ and calculates and sends $g_b = g^b \bmod P$ to Alice.

Alice and Bob can now create the shared secret $g^{ab} \bmod P$. Alice calculates this as $g_b^a \bmod P$ and Bob calculates this as $g_a^b \bmod P$. Note that all of these terms can be calculated in time logarithmic in a and b through repeated squaring.

An adversary, however, would see only $g^a \bmod P$ and $g^b \bmod P$, and would also know g and P . How could $g^{ab} \bmod P$ be determined from this? One method is to solve what is called the *discrete logarithm problem*. This is the problem of calculating a if you know P , g , and $g^a \bmod P$. For large P this appears to be a computationally intractable problem. It is not known for certain whether there is a faster way for the adversary to calculate $g^{ab} \bmod P$ than computing discrete logarithms—this is called the *Diffie-Hellman problem*—but at present no better method is known.

It is not obvious how to find primitive elements in general, but it is much easier if, as is usually the case, the prime P has been constructed so as to ensure that the factorization of $P - 1$ is known. For instance, if P is of the form $2Q + 1$, where Q is also a prime (such numbers are called *Sophie Germain primes*), then it can be

shown that for any a , exactly one of a and $-a$ has the property that its Q th power is congruent to $-1 \bmod P$, and this one is a primitive element. In practice, one can find such primes by a process of trial and error: for example, one can choose a number Q randomly and use randomized primality tests to see whether Q and $2Q + 1$ are prime. Assuming that, as everyone believes, such pairs occur with the “expected” frequency, the probability of finding one on any given attempt is large enough for this approach to be feasible.

5.3 Other Groups

The Diffie-Hellman protocol can be expressed in the language of GROUP THEORY [I.3 §2.1]. Suppose we have a group G and some element $g \in G$. We will require the group to be Abelian and will use “+” to denote the group operation. (In the examples so far, the groups under consideration were multiplicative groups consisting of elements coprime to some integer N , so by using additive notation we are taking a “logarithmic” perspective.)

To execute the protocol Alice computes some private integer a and computes and sends ag to Bob. Note that Alice can compute this sum of a elements of G in time of order logarithmic in a by successive doubling and adding. (In the multiplicative groups considered earlier, “doubling” is squaring, “adding” is multiplying, and “multiplying by a ” is raising to the power a .)

Similarly, Bob computes a private integer b and computes and sends bg to Alice.

Both Alice and Bob can calculate the shared value abg . An adversary will know only G , g , ag , and bg .

The question is: which groups can be used in practical cryptographic systems? The critical property is that the discrete logarithm problem in G must be hard; in other words, given G , g , and ag it should be a hard problem to determine a .

One type of group that has aroused interest for cryptographic purposes is the additive group generated by points on an ELLIPTIC CURVE [III.21]. An elliptic curve has an equation of the form

$$y^2 = x^3 + ax + b.$$

It is an interesting exercise to sketch this curve over the real numbers—the shape depends upon how many times the curve

$$y = x^3 + ax + b$$

crosses the x -axis.

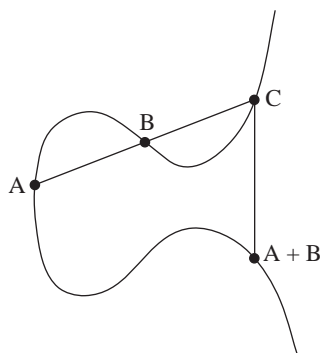


Figure 3 Addition of points on an elliptic curve.

It is possible to define an “addition rule” (often called a *group law*) on the points of this curve, as follows. Given two points A and B on the curve, the straight line joining them must meet the curve in a third point, C say. This is because a straight line must meet a cubic in three places precisely. Define $A + B$ to be the mirror image of C in the x -axis (see figure 3).

It is obvious that $A + B = B + A$ from this definition. What is rather more surprising is that the associative law holds. That is, for any three points A, B, and C we have $((A + B) + C) = (A + (B + C))$. There are some deep reasons why this is true, but of course it can be verified by just doing the algebra.

To use this for cryptography the group is formed from the set of points on an elliptic curve defined over a finite field. The graphical image for the sum of two points is no longer valid, but the algebraic definition still holds, so addition still obeys the associative law. We need to add one further point to the set of points on the curve to function as the zero of the group: this is the “point at infinity” on the curve.

For optimal security it turns out to be best to find a curve defined over \mathbb{F}_p for which the number of elements in the group is a prime number. In fact it is guaranteed—by a deep result on the theory of elliptic curves—that the number of points on a curve defined over \mathbb{F}_p will lie between $p + 1 - 2\sqrt{p}$ and $p + 1 + 2\sqrt{p}$. (See THE WEIL CONJECTURES [V.38].)

The reason this group is used is that for general curves the discrete logarithm problem appears to be particularly hard. If the group has n elements and if we are given group elements g and ag , then the number of steps needed to determine a , by the best algorithms that are currently known, is around \sqrt{n} . Since there is a so-called birthday attack that allows one to solve this problem in *any* group with n elements in around \sqrt{n}

computational steps, this means that the problem for elliptic curve groups is as hard as it can be. Therefore, whatever level of security you require, the public key is as short as it can be. This is important when there are constraints on the number of bits that can be sent as it allows the protocol to be executed in the minimum possible time.

6 Digital Signatures

As well as secure transmission of data, there is another very useful capability that is provided by public-key cryptography. That is the concept of a *digital signature*. A digital signature is a string of symbols that an author attaches to the end of a message that certifies the authenticity of the message. In other words, it proves that the message was written by the attested author and that it has not been modified. Once the necessary frameworks are in place, this opens up the possibility of much legal business being conducted online.

There are a number of ways that public-key methods can be used to create digital signatures. The one based on the RSA system is perhaps the simplest. Suppose Alice wants to sign documents. Just as she does for encryption, she generates two large prime numbers P and Q and calculates her public modulus $N = PQ$ and her public exponent e . She also generates her private key—the deciphering exponent d with the property that $x^{de} \equiv x \pmod{N}$ for any x . She will use the same parameters both for encryption and for the creation of digital signatures.

Alice can assume that the recipients of her signed messages know her N and e values. In practice she may have these values themselves signed and certified by a trusted authority or organization that the prospective recipient of a signed message will recognize.

One other component of this system is an object called a *one-way hash function*, which takes as its input the message to be signed, which may be rather long, and outputs a number between 1 and $N - 1$. The important property that a hash function must have is that for any value y between 1 and N it is computationally hard to construct a message x that hashes to that value. This is similar to a one-way function except that we are no longer assuming that for each y there is exactly one x that maps to y . However, the hash function should ideally also be *collision free*, which means that, even though there are many pairs of messages that hash to the same value, it is not easy to find any. Such hash functions need to be carefully designed, but there

are some recognized standard hash functions (two of which are called MD5 and SHA-1). Suppose that x is the message to be signed, and let X be the output when you apply the hash function to x . The digital signature that Alice appends to the message is $Y = X^d \bmod N$.

Observe that anyone in possession of Alice's public key can verify the signature by following these steps. First, calculate the hashed value X of the message x , which is possible because the hash function is made public. Next, compute $Z = Y^e \bmod N$, which can be done because the parameters N and e are also public. Finally, verify that X equals Z . In order to fake such a signature, you have to find Y with the property that $Y^e \equiv X \bmod N$. That is, you must know how to calculate X^d , which is computationally intractable if you do not already know d .

It is also possible to construct digital signatures using a public key based on discrete logarithms (Diffie-Hellman type) rather than on factorization (RSA type). The U.S. standards body has published such a proposal: the *Digital Signature Standard* (1994).

7 Some Current Research Topics

Cryptography remains an active and fascinating area for research—there are undoubtedly more results and ideas to be discovered. For a good overview of current activity one should look at recent proceedings of the main conferences, such as Crypto, Eurocrypt, or Asiacrypt (these are published in the Springer series Lecture Notes in Computer Science). The comprehensive book on cryptography by Menezes, van Oorschott, and Vanstone (1996) is a good way to get up to speed on present theory. In this final section I outline just a few of the directions in which the subject is moving.

7.1 New Public-Key Methods

One important area of investigation is the search for new public-key methods and signature schemes. Recently some interesting new ideas have come from the use of *pairings* on elliptic curves (Boneh and Franklin 2001). These are maps w from pairs of points on the curve to either the finite field over which the curve is defined or an extension field.

A pairing w is *bilinear*, in the sense that $w(A+B, C) = w(A, C)w(B, C)$ and $w(A, B+C) = w(A, B)w(A, C)$, where addition is the group operation defined on points of the curve and multiplication takes place in the field.

One way that such a map can be used is to create an "identity-based cryptosystem." Here, a user's identity

serves as his or her public key, which eliminates the need for directories or other public-key infrastructure in order to store and propagate public keys.

In such a system, a central authority decides upon a curve, a pairing map w , and a hash function that maps identities to points on the curve. All of this is made public, but there is also a secret parameter, an integer x .

Suppose that the hash function maps Alice's identity to the point A on the curve. The authority calculates Alice's private key xA and issues it to her when she registers, after making appropriate checks on her identity. Similarly, Bob would receive his private key xB , where B is the point on the curve corresponding to his identity.

Alice and Bob are now able to communicate without any initial key exchange, using the common key $w(xA, B) = w(A, xB)$. The important point is that unlike other public-key systems this can be done without any need to share public keys.

7.2 Communication Protocols

A second area of activity is the study of proposed protocols, especially those likely to become international standards. When public-key methods are to be used in practical communication the sequence of bits to be transmitted needs to be clearly defined, so that both communicating parties understand the same thing by each bit sent. For example, if an n -bit number is transmitted, are the bits transmitted in increasing or decreasing order of significance? The rules or protocols are often enshrined in public standards, and it is important that they do not introduce any weakness into the system.

An example of the sort of weakness that can be introduced in this way is one discovered by Coppersmith (1997) in a seminal paper. He showed that in a low-exponent RSA system (for example, one with encryption exponent equal to 17) a weakness arises if too many of the bits of the number that is to be enciphered are set to publicly known values. This is something that is natural to want to do, if, as is often the case, a large public-key modulus is being used to transmit a much shorter communication key. As a result of Coppersmith's discovery such fields are nowadays usually padded out before they are encrypted, with bits that vary unpredictably.

7.3 Control of Information

Using public-key methods, one can control very precisely how information is released, shared, or generated. Research in this area is usually focused on finding elegant and efficient ways of achieving different sorts of control in a variety of situations. As a simple example, we might want to create a secret that is shared between N people in such a way that if any K people combine their share (where $K < N$) they can reconstruct the secret, but no information can be gained about the secret by any smaller number than K collaborating.

Another example of this type of control is a protocol that allows two participants to create an RSA modulus (a product of two primes) in such a way that neither participant gets to know the primes that were used to produce the modulus. To decipher a message enciphered under this modulus the two participants have to collaborate—neither can achieve this on their own (Cocks 1997).

A third and more amusing example is a protocol that allows Alice and Bob to replicate tossing a coin, but to do it over the telephone. Obviously, it would not be satisfactory for Alice to toss the coin and for Bob to make the call “heads” or “tails”—for how does Bob know that Alice is telling the truth about how the coin actually fell? This problem turns out to have a simple solution. Alice and Bob choose large random sequences. Alice then appends either a 1 or a 0 to her sequence and Bob does the same for his. Alice’s extra bit represents the outcome of the coin toss, and Bob’s represents his guess. Next, they send one-way hashes of their sequences (with the extra bits appended). At this point, because of the nature of one-way hashes, neither has any idea what the other’s sequence is, so, for example, if Alice reveals her hashed sequence first, Bob cannot use this information to increase his chance of guessing correctly. Alice and Bob then exchange the unhashed sequences to see whether Bob’s guess was correct. If either does not trust the other, they can hash the other’s sequence to check that it really does give the right answer. Since it is hard to find a different sequence that gives the right answer, they can each be confident that the other has not cheated. More complicated protocols of this type have been designed—it is even possible to play poker remotely in this way.

Further Reading

Boneh, D., and M. Franklin. 2001. Identity-based encryption from the Weil pairing. In *Advances in Cryptology*—

- CRYPTO 2001*. Lecture Notes in Computer Science, volume 2139, pp. 213–29. New York: Springer.
- Cocks, C. 1997. *Split Knowledge Generation of RSA Parameters*. *Cryptography and Coding*. Lecture Notes in Computer Science, volume 1355, pp. 89–95. New York: Springer.
- Coppersmith, D. 1997. Small solutions to polynomial equations, and low exponent RSA vulnerabilities. *Journal of Cryptology* 10(4):233–60.
- Daeman, J., and V. Rijmen. 2002. *The Design of Rijndael*. AES—The Advanced Encryption Standard Series. New York: Springer.
- Data Encryption Standard. 1999. Federal Information Processing Standards Publications, number 46-3.
- Diffie, W., and M. Hellman. 1976. New directions in cryptography. *IEEE Transactions on Information Theory* 22(6): 644–54.
- Digital Signature Standard. 1994. Federal Information Processing Standards Publications, number 186.
- Lenstra, A., and H. Lenstra Jr. 1993. *The Development of the Number Field Sieve*. Lecture Notes in Mathematics, volume 1554. New York: Springer.
- Lenstra Jr., H. 1987. Factoring integers with elliptic curves. *Annals of Mathematics* 126:649–73.
- Massey, J. 1969. Shift-register synthesis and BCH decoding. *IEEE Transactions on Information Theory* 15:122–27.
- Menezes, A., P. van Oorschott, and S. Vanstone. 1996. *Applied Cryptography*. Boca Raton, FL: CRC Press.
- Rivest, R., A. Shamir, and L. Adleman. 1978. A method for obtaining digital signatures and public-key cryptosystems. *Communications of the Association for Computing Machinery* 21(2):120–26.
- Silverman, R. 1987. The multiple polynomial quadratic sieve. *Mathematics of Computation* 48:329–39.
- Singh, S. 1999. *The Code Book*. London: Fourth Estate.

VII.8 Mathematics and Economic Reasoning

Partha Dasgupta

1 Two Girls

1.1 Becky’s World

Becky, who is ten years old, lives with her parents and an older brother Sam in a suburban town in America’s Midwest. Becky’s father works in a law firm specializing in small business enterprises. Depending on the firm’s profits, his annual income varies somewhat, but it is rarely below \$145 000. Becky’s parents met in college. For a few years her mother worked in publishing, but when Sam was born she decided to concentrate on raising a family. Now that both Becky and Sam

Terri: Tim thinks that what is here is clearer than any of the alternatives and that this use of ‘they’ is “so convenient that it’s wrong to struggle to avoid it”. I tend to agree but obviously you should let me know if you don’t.

attend school, she does voluntary work in local education. The family live in a two-storey house. It has four bedrooms, two bathrooms upstairs and a toilet downstairs, a large drawing-cum-dining room, a modern kitchen, and a family room in the basement. There is a small plot of land in the rear, which the family use for leisure activities.

Although their property is partially mortgaged, Becky's parents own stocks and bonds and have a savings account in the local branch of a national bank. Becky's father and his firm jointly contribute to his retirement pension. He also makes monthly payments into a scheme with the bank that will cover college education for Becky and Sam. The family's assets and their lives are insured. Becky's parents often remark that, federal taxes being high, they have to be careful with money; and they are. Nevertheless, they own two cars, the children attend camp each summer, and the family take a vacation together once camp is over. Becky's parents also remark that her generation will be much more prosperous than they. Becky wants to save the environment and insists on biking to school each day. Her ambition is to become a doctor.

1.2 Desta's World

Desta, who is about ten years old, lives with her parents and five siblings in a village in subtropical, southwest Ethiopia. The family live in a two-room, grass-roofed mud hut. Desta's father grows maize and *tef* on half a hectare of land that the government has awarded him. Desta's older brother helps him to farm the land and care for the household's livestock: a cow, a goat, and a few chickens. The small quantity of *tef* produced is sold so as to raise cash income, but the maize is largely consumed by the household as a staple. Desta's mother works a small plot next to their cottage, growing cabbage, onions, and *enset* (a year-round root crop that also serves as a staple). In order to supplement household income, she brews a local drink made from maize. As she is also responsible for cooking, cleaning, and minding the infants, her work day usually lasts fourteen hours. Despite the long hours, it would not be possible for her to complete the tasks on her own. (As the ingredients are all raw, cooking alone takes five hours or more.) So Desta and her older sister help their mother with household chores and mind their younger siblings. Although a younger brother attends the local school, neither Desta nor her older sister has ever been enrolled there. Her parents can neither read nor write, but they are numerate.

Desta's home has no electricity or running water. Around where they live, sources of water, land for grazing cattle, and the woodlands are communal property. They are shared by people in Desta's village; but the villagers do not allow outsiders to make use of them. Each day Desta's mother and the girls fetch water, collect fuelwood, and pick berries and herbs from the local commons. Desta's mother frequently observes that the time and effort needed to collect their daily needs has increased over the years.

There is no financial institution nearby to offer either credit or insurance. As funerals are expensive occasions, Desta's father long ago joined a community insurance scheme (*iddir*) to which he contributes monthly. When Desta's father purchased the cow they now own, he used the entire cash he had accumulated and stored at home, but had to supplement that with funds borrowed from kinfolk, with a promise to repay the debt when he had the ability to do so. In turn, when they are in need, his kinfolk come to him for a loan, which he supplies if he is able to. Desta's father says that such patterns of reciprocity he and those close to him practice are part of their culture, reflecting their norms of social conduct. He also says that his sons are his main assets, as they are the ones who will look after him and Desta's mother in their old age.

Economic statisticians estimate that, adjusting for differences in the cost of living between Ethiopia and the United States, Desta's family income is about \$5000 per year, of which \$1000 is attributable to the products they draw from the local commons. However, as rainfall varies from year to year, Desta's family income fluctuates widely. In bad years, the grain they store at home gets depleted well before the next harvest. Food is then so scarce that they all grow weaker, the younger children especially so. It is only after harvest that they regain their weight and strength. Periodic hunger and illnesses have meant that Desta and her siblings are somewhat stunted. Over the years Desta's parents have lost two children in their infancy, stricken by malaria in one case and diarrhea in the other. There have also been several miscarriages.

Desta knows that she will be married (in all likelihood to a farmer, like her father) when she reaches eighteen and will then live on her husband's land in a neighboring village. She expects her life to be similar to that of her mother.

2 The Economist's Agenda

That the lives people are able to construct differ enormously across the globe is a commonplace. In our age of travel, it is even a common sight. That Becky and Desta face widely different futures is also something we have come to expect, perhaps also to accept. Nevertheless, it may not be out of turn to imagine that the two girls are intrinsically very similar: they both enjoy eating, playing, and gossiping; they are close to their families; they like pretty things to wear; and they both have the capacity to be disappointed, get annoyed, be happy. Their parents are also alike. They are knowledgeable about the ways of their worlds. They also care about their families, finding ingenious ways to meet the recurring problem of producing income and allocating resources among family members—over time and allowing for unexpected contingencies. So, a promising route for exploring the underlying causes behind their vastly different conditions of life would be to begin by observing that the constraints the families face are very different: that in some sense Desta's family are far more restricted in what they are able to be and do than Becky's.

Economics in large measure tries to uncover the processes that influence how people's lives come to be what they are. The context may be a household, a village, a district, a state, a country, or the whole world. In its remaining measure, the discipline tries to identify ways to influence those very processes so as to improve the prospects of those who are hugely constrained in what they can be and do. *Modern* economics, by which I mean the style of economics taught and practiced in today's graduate schools, does the exercises from the ground up: from individuals, through the household, village, district, state, country, to the whole world. In varying degrees the millions of individual decisions shape the eventualities people all face; as both theory and evidence tell us that there are enormous numbers of unintended consequences of what we all do. But there is also a feedback, in that those consequences go on to shape what people subsequently can do and choose to do. For example, when Becky's family drive their cars or use electricity, or when Desta's family create compost or burn wood for cooking, they contribute to global carbon emissions. Their contributions are no doubt negligible, but the millions of such tiny contributions cumulatively sum to a sizable amount, having consequences that people everywhere are likely to experience in different ways.

To understand Becky's and Desta's lives, we need first of all to identify the prospects they face for transforming goods and services into further goods and services—now and in the future, under various contingencies. Second, we need to uncover the character of their choices and the pathways by which the choices made by millions of households like Becky's and Desta's go to produce the prospects they all face. Third, and relatedly, we need to uncover the pathways by which the families came to inherit their current circumstances.

The last of these is the stuff of economic history. In studying history we could, should we feel bold, take the long view—from about the time agriculture came to be settled practice in the Fertile Crescent (roughly, Anatolia) some eleven thousand years ago—and try to explain why the many innovations and practices that have cumulatively contributed to the making of Becky's world either did not reach or did not take hold in Desta's part of the world. (Diamond (1997) is an enquiry into this set of questions.) If we wanted a sharper account, we could study, say, the past six hundred years and ask how it is that, instead of the several regions in Eurasia that were economically promising in about 1400 C.E., it was the unlikely northern Europe that made it and helped to create Becky's world, even while bypassing Desta's. (Landes (1998) is an inquiry into that question. Fogel (2004) explores the pathways by which Europe during the past three hundred years has escaped permanent hunger.) As modern economics is largely concerned with the first two sets of enquiries, this article focuses on them. However, the methods that today's economic historians deploy to answer their questions are not dissimilar to the ones I describe below to study contemporary lives. The methods involve studying individual and collective choices in terms of *maximization exercises*. The predictions of the theories are then tested by studying data relating to actual behavior. Even the ethical foundations of national economic policies involve maximization exercises: the maximization of social well-being subject to constraints. (The treatise that codified this approach to economic reasoning was Samuelson (1947).)

3 The Household Maximization Problem

Both Becky's and Desta's households are micro-economies. Each subscribes to particular arrangements over who does what and when, recognizing that it faces constraints on what its members are capable of doing.

We imagine that both sets of parents have their respective families' well-being in mind and want to do as well as they can to protect and promote it.¹ Of course, both Becky's and Desta's parents would have a wider notion of what constitutes their families than I have allowed here. Maintaining ties with kinfolk would be an important aspect of their lives, a matter I return to later. One also imagines that Becky's and Desta's parents are interested in their future grandchildren's well-being. But as they recognize that their children will in turn care about *their* children, they are right to conclude by recursion that doing the best for their children amounts to doing the best for their grandchildren, for their great grandchildren, and so on, down the generations.

Personal well-being is made up of a variety of constituents: health, relationships, place in society, and satisfaction at work are but four. Economists and psychologists have identified ways to represent well-being as a numerical measure. To say that someone's well-being is greater in situation Y than in situation Z is to say that her well-being measure is numerically higher in Y than in Z . A family's well-being is an aggregate of its members' well-beings. As goods and services are among the determinants of well-being (some important examples are food, shelter, clothing, and medical care), the problem that both Becky's and Desta's parents face is to determine, from among those allocations of goods and services that are feasible, the ones that are best for their households. However, both pairs of parents care not only about today, but also about the future. Moreover, the future is uncertain. So when the parents think about which goods and services their households should consume, they are concerned not just with the goods and services themselves, but also with when they will be consumed (food today, food tomorrow, and so on) and what will happen in the case of various contingencies (food the day after tomorrow if rainfall turns out to be bad tomorrow, and so forth). Implicitly or explicitly, both sets of parents convert their experience and knowledge into probabilistic judgments. Some of the probabilities they attach to contingencies are no doubt very subjective, but others, such as their predictions about the weather, are arrived at from long experience.

1. As suggested by McElroy and Horney in 1981, a realistic alternative would be to suppose that household decisions are reached by negotiation between the various parties (see Dasgupta 1993, chapter 11). Qualitatively, nothing much is lost in my assuming optimizing households here.

In subsequent sections we shall study the way in which Becky's and Desta's parents allocate goods and services across time and contingencies. But here we shall keep the exposition simple and consider a model that is *static* and *deterministic*. That is, we shall pretend that the people live in a timeless world, and that they are completely certain about all the information they need in order to make their decisions.

Suppose that a certain household has N members, whom we label $1, 2, \dots, N$. Let us think about how we can appropriately model the well-being of household member i . As has already been mentioned, well-being is taken to be a real number that depends in some way on the goods and services consumed and supplied by i . It is traditional to divide goods and services into those *consumed* and those *supplied*, and to use positive numbers to represent quantities of the former and negative numbers for the latter. Imagine now that there are M commodities in all. Let $Y_i(j)$ represent the quantity of the j th commodity that is consumed or supplied by i . By our convention, $Y_i(j) > 0$ if j is consumed by i (e.g., food eaten or clothing worn) and $Y_i(j) < 0$ if j is supplied by i (e.g., labor). Now consider the vector $\mathbf{Y}_i = (Y_i(1), \dots, Y_i(M))$. It denotes the quantities of all the goods and services consumed or supplied by i . \mathbf{Y}_i is a point in \mathbb{R}^M —the Euclidean space of M dimensions. We now let $U_i(\mathbf{Y}_i)$ denote i 's well-being. Let us assume that supplying goods and services decreases i 's well-being, while consuming them increases it. Because the goods that are supplied by i are measured as negative quantities, we can justifiably assume that $U_i(\mathbf{Y}_i)$ increases as any of its components \mathbf{Y}_i increases.

The next step is to generalize the model to a household. The individual well-beings of the members of the household can be collected together so that they themselves form an N -dimensional vector, $(U_1(\mathbf{Y}_1), \dots, U_N(\mathbf{Y}_N))$. The household's well-being is dependent in some way on this vector. That is, we say that the well-being of the household is $W(U_1(\mathbf{Y}_1), \dots, U_N(\mathbf{Y}_N))$, for some function W . (Utilitarian philosophers have argued that W is simply the *sum* of the U_i .) We also make the natural assumption that W is an increasing function of each U_i (which is certainly the case if W is the sum of the U_i).

Let \mathbf{Y} denote the sequence $(\mathbf{Y}_1, \dots, \mathbf{Y}_N)$. \mathbf{Y} is a point in the NM -dimensional Euclidean space \mathbb{R}^{NM} . It can also be thought of as the matrix you obtain if you make a table of the amounts of each commodity consumed or supplied by each member of the household. Now, it is clear that not every \mathbf{Y} in \mathbb{R}^{NM} can actually occur:

after all, the total amount of any given commodity (in the whole world, say) is finite. So we assume that Y belongs to a certain set J , which we regard as the set of all *potentially feasible* values of Y . Within J we identify a smaller set, F , of “actually feasible” values of Y . This is the set of values of Y from which the household could in principle choose. It is smaller than J because of constraints that the household faces, such as the maximum amount of income it can earn. F is the household’s *feasible set*.² The decision faced by a household is to choose Y from the feasible set F so as to maximize its well-being $W(U_1(Y_1), \dots, U_N(Y_N))$. This is called *the household maximization problem*.

It is reasonable, and mathematically convenient, to assume that the sets J and F are both closed and bounded subsets of \mathbb{R}^{NM} , and that the well-being function W is continuous. Since every continuous function on a closed bounded set has a maximum, it follows that the household maximization problem has a solution. If, in addition, W is differentiable, the theory of *non-linear programming* can be used to identify the optimality conditions the household’s choice must satisfy. If F is a convex set and W is a concave function of Y , those conditions are both necessary and sufficient. The LAGRANGE MULTIPLIERS [III.66] associated with F can be interpreted as *notional prices*: they reflect the worth to the household of slightly relaxing the constraints.

Let us conduct an exercise to test the power of the modern economist’s way of studying choice. First, let us assume that W is a *symmetric* and *concave* function of the individual well-beings U_i (as would be the case if W were the sum of the U_i). The symmetry assumption means that if two individuals exchange their well-beings, then W is unchanged; and concavity means, roughly speaking, that, other things being equal, as a U_i increases, the rate of increase of W does not rise. Let us suppose in addition that the household members are identical: that is, let us set all the functions U_i to be equal to a single function, U , say. Assume also that U is a strictly concave function of the Y_i , which means that the rate of increase of well-being declines as consumption increases. Finally, assume that the feasible set F is nonempty, convex, and symmetric. (Symmetry means that if some Y is feasible, and the vector Z is the same as Y except that the consumptions of a pair of individuals in the household have been exchanged, then Z is also feasible.) From these assumptions it can be shown that members of the household would be treated

equally: that is, W is maximized when they all receive the same bundle of goods and services.

At low levels of consumption, however, the hypothesis that the function U is concave is unreasonable. To see why, we should note that, typically, 60–75% of the daily energy intake of someone in nutritional balance goes toward maintenance, while the remaining 25–40% is expended in discretionary activities (work and leisure). The 60–75% is rather like a fixed cost: over the long run a person needs it as a minimum no matter what he or she does. The simplest way to uncover the implications of such fixed costs is to continue to suppose that F is convex (which is the case, for example, if there is a fixed quantity of food for allocation among members of the household), but that U is a strictly convex function at low intakes of food and a strictly concave function thereafter. It is not hard to show that a poor household in such a world will maximize its well-being by allocating food unequally among its members, while a rich household can afford the luxury of equal treatment and will choose to distribute food equally. Suppose, to take a very stylized example, that energy requirement for daily maintenance is 1500 kcal and that a household of four can obtain at most 5000 kcal for consumption. Then equal sharing would mean that no one would have sufficient energy for any work, so it is better to share the food unequally. On the other hand, if the household is able to obtain more than 6000 kcal, it can share the food equally without jeopardizing its future.

There are empirical correlates of this finding. When food is very scarce, the younger and weaker members of Desta’s household are given less to eat than the others, even after allowance is made for differences in their ages. In good times, though, Desta’s parents can afford to be egalitarian. In contrast, Becky’s household can always afford enough food. Her parents therefore allocate food equally every day.

4 Social Equilibrium

Household transactions in Becky’s world are carried out mostly in markets. The terms of trade are the quoted market prices. In developing a mathematical construction of social outcomes, I continue to imagine, for simplicity, a static, deterministic world. Let $P (\geq 0)$ be the vector of market prices and let $M (\geq 0)$ be the vector of a household’s endowments of goods and services. (That is, for each commodity j , $P(j)$ is the price of j and $M(j)$ is the amount of j that the household already

2. Presently we will see why we need to distinguish J from F , rather than looking just at F .

has.) Recalling our convention that goods consumed are of positive sign and goods supplied are of negative sign, define $X = \sum Y_i$. (Thus, $X(j) = \sum Y_i(j)$ is the total amount of commodity j that is consumed by the household.) Then $P \cdot X$ is the total price of goods consumed by the household, minus the total price of goods supplied, and $P \cdot M$ is the total value of its endowments. The feasible set F is the set of household choices Y that satisfy the “budget” constraint $P \cdot (X - M) \leq 0$.

The income that Becky’s household earns from the assets it supplies to the market is determined by market prices (Becky’s father’s salary, interest rates on bank deposits, returns on shares owned). Those prices in turn depend on the size and distribution of household endowments of goods and services and on household needs and preferences. They depend too on the ability and willingness of institutions, such as private firms and the government, to make use of the rights they in turn have been awarded. These functional relationships explain why Becky’s father’s skills as a lawyer (itself an asset, termed “human capital” by economists) would not be worth much in Desta’s village, even though they are much valued in the United States. In fact, it was a firm belief that lawyers would continue to prove valuable in the United States that encouraged Becky’s father to *be* a lawyer.

Although Desta’s household does operate in markets (when her father sells *tef* or her mother sells the liquor she has brewed), it undertakes many transactions directly with nature; in the local commons and in farming, and in nonmarket relationships with others in the village. Therefore the F that Desta’s household faces is not defined simply by a linear budget inequality, as in the idealized model we have constructed to display Becky’s world, but also reflects the constraints that nature imposes, such as soil productivity and rainfall, the assets it has access to, and the terms and conditions involving transactions with others in the village via nonmarket relationships, a matter I come to later. The constraints imposed by nature are felt by Becky’s household too, but through market prices. For example, should a drought lead to a fall in world cereal production, it would become noticeable to Becky’s household through the high price of cereal. Desta’s household, in contrast, would notice it directly from the reduced harvest from their field.

Desta’s household assets include the family home, livestock, agricultural implements, and their half hectare of land. The skills Desta’s family members have

accumulated in farming, managing livestock, and collecting resources from the local commons are part of their human capital. Those skills do not command much return in the global marketplace, but they do shape the household’s feasible set F and are vital to the family’s well-being. Desta’s parents learned those skills from their parents and grandparents, just as Desta and her siblings have learned them from their parents and grandparents. Desta’s family can also be said to own a portion of the local commons: in effect, her household shares its ownership with others in the village. Difficulties in reaching and enforcing agreement with neighbors over the use of the local commons are less severe than they are in the case of global commons, such as the atmosphere as a sink for carbon emissions. This is not only because the required negotiations involve far fewer people when the commons are local, but also because there is likely to be greater congruence of opinions and interests among the users. It helps too that the parties are able to observe whether the agreements they made over the use of local commons are being kept. (See below in our discussion of insurance arrangements in Desta’s world.)

Thus, the choices other people make affect the choices that are available to individuals, which results in feedback. In a market economy, the feedback is in large part transmitted in prices. In nonmarket economies the feedback is transmitted through the terms in which households are able to negotiate with one another.

Let us try to model this situation mathematically. We start by imagining an economy of H households. For ease of exposition, I shall suppose that a household’s well-being can be expressed directly in terms of its aggregate consumption of goods and services, disregarding how this consumption is distributed among the individual members. Let X_h denote the consumption vector in household h (with the usual sign convention), let J_h be the set of potentially feasible vectors X_h , and let $W_h(X_h)$ be h ’s well-being.

Within h ’s potentially feasible set J_h of consumption vectors lies the actual feasible set F_h . In order to model the feedback we shall explicitly recognize that F_h depends on the consumptions of other households. That is, it is a function of the sequence $(X_1, \dots, X_{h-1}, X_{h+1}, \dots, X_H)$. To save space, we shall denote this sequence, which consists of every household’s consumption vector except h ’s, by X_{-h} . Formally, F_h is a function (sometimes called a “correspondence”) that takes objects of the form X_{-h} to subsets of J_h .

Household h 's economic problem is to choose its consumption X_h from its feasible set $F_h(X_{-h})$ in such a way as to maximize its well-being $W_h(X_h)$. The optimum choice depends on h 's beliefs about X_{-h} and the correspondence $F_h(X_{-h})$.

Meanwhile, all other households are making similar calculations. How can we unravel the feedbacks? One way would be to ask people to disclose their beliefs about the feedbacks. Fortunately, economists avoid that route. So as to anchor their investigation, economists study *equilibrium* beliefs; that is, beliefs that are self-confirming. The idea is to identify states of affairs where the choices people make on the basis of their beliefs about the feedbacks are precisely those that give rise to those very feedbacks. We call any such state of affairs a *social equilibrium*. Formally, a sequence (X_1^*, \dots, X_H^*) of household choices is called a *social equilibrium* if, for every h , the choice X_h^* of household h maximizes the well-being $W_h(X_h)$ over all choices of X_h in its feasible set $F_h(X_{-h}^*)$.

This raises an obvious question: does a social equilibrium exist? Classic papers by Nash in 1950 and Debreu in 1952 showed that, under a fairly general set of conditions, it always does. Here is a set of conditions that Debreu identified. Assume that each well-being function W_h is continuous and *quasi-concave* (which means that for any potentially feasible choice X'_h in J_h , the set of X_h in J_h for which $W_h(X_h)$ is greater than or equal to $W_h(X'_h)$ is convex). Assume also that for every household h , the feasible set F_h (recall that this is a subset of J_h) is nonempty, compact, and convex, and depends continuously on the choices X_{-h} made by other households. The proof that under the above conditions a social equilibrium always exists is a relatively straightforward use of THE KAKUTANI FIXED-POINT THEOREM [V.13 §2], which is itself a generalization of Brouwer's fixed-point theorem. Alternative sets of sufficient conditions for the existence of social equilibria (which allow the feasible set $F_h(X_{-h})$ to be nonconvex) have been explored in recent years.

In Becky's world, a social equilibrium is called a *market equilibrium*. A market equilibrium is a price vector $P^* (\geq 0)$ and a consumption vector X_h^* for each household h , such that X_h^* maximizes $W_h(X_h)$ subject to the budget constraint $P^* \cdot (X_h - M_h) \leq 0$, and such that the demands for goods and services across households are feasible (i.e., $\sum (X_h - M_h) \leq 0$). That market equilibria are social equilibria, in the sense in which we have defined the latter term here, was demonstrated by

Arrow and Debreu in 1954. Debreu (1959) is the definitive treatise on market equilibria. In that book, Debreu followed the leads of Erik Lindahl and Kenneth J. Arrow, by distinguishing goods and services not only in terms of their physical characteristics, but also in terms of the date and contingency in which they appear. Later in this article we shall expand the commodity space in that way to study savings and insurance decisions in both Becky's and Desta's worlds.

One cannot automatically assume that a social equilibrium is just or collectively good. Moreover, except for the most artificial examples, social equilibrium is not unique—which means that a study of equilibria per se leaves open the question of which social equilibrium we should expect to observe. In order to probe that question, economists study disequilibrium behavior and analyze the stability properties of the resulting dynamic processes. The basic idea is to hypothesize about the way people form beliefs about the way the world works, track the consequences of those patterns of learning, and check them against data. It is reasonable to limit such a study by considering only those learning processes that converge to a social equilibrium in stationary environments. Initial beliefs would then dictate which equilibrium is reached in the long run (see, for example, Evans and Honkapohja 2000). Since the study of disequilibria would lengthen this article greatly, we shall continue to study social equilibria here.

5 Public Policy

Economists distinguish between what they call *private goods* and *public goods*. For many goods, consumption is rivalrous: if you consume a bit more from a given supply of such a good (e.g., food), others have that much less to consume. These are private goods. The way to assess their consumption throughout the economy is to add up the amounts consumed by all individual households; which is what we did in the previous section when arriving at the notion of a social equilibrium. Not all goods are like that, however. For example, the extent of national security on offer to you is the same as that on offer to all households in your country. In a just society the law has that same property, as has the state: not only is consumption *not* rivalrous, but in addition, no one can be prevented from availing himself or herself of the entire amount available in the economy. Public goods are goods of this second kind. One models the quantity of a public good as a number G , and the quantity G_h consumed by each household h is deemed to

equal G . An example of a public good that has a global coverage is the Earth's atmosphere: the whole world benefits from it jointly.

If the supply of public goods is left to private individuals, then problems arise. For example, even though everyone in a city would benefit from a cleaner, healthier environment, individuals have a strong incentive to free-ride on others when it comes to paying for that cleaner environment. Samuelson showed in 1954 that such a situation resembles the prisoner's dilemma: each party has a strategy that is best for him/her, regardless of what strategies the other parties choose, even though there is another set of strategies, one per party, that is better for everybody. Under such circumstances, one usually needs public measures, such as taxes and subsidies, in order for it to be in the interest of private individuals to act in a way that implements the collectively preferred outcomes. In other words, the dilemma can be expected to be resolved effectively not by markets but by politics. It is widely accepted in political theory that government should be charged with imposing taxes, subsidies, and transfers, and should be engaged in supplying public goods. The government is also the natural agency to supply infrastructure, such as roads, ports, and electrical cables, requiring as they do investments that are huge in comparison with individual incomes. We shall now extend our earlier model to include public goods and infrastructure, so that we can study the government's economic task.

Let us assume that social well-being is a numerical aggregate of household well-beings. Thus, if V is social well-being, we write it as $V(W_1, \dots, W_H)$. It is natural to postulate that V increases as any W_h increases. (One example of such a function V is the one prescribed by utilitarian philosophy, namely, $W_1 + \dots + W_H$.) The government chooses what quantities to supply of the various public goods and infrastructure commodities. These numbers can be modeled by two vectors, which we will call G and I , respectively. The government also chooses to impose on each household h certain transfers T_h of goods and services (for example, providing health care and charging income tax). Let us write T for the sequence (T_1, \dots, T_H) . Whether or not a particular choice of vectors G and I is actually feasible for the government will depend on T , so we define K_T to be the set of feasible pairs of vectors (G, I) , given the choice of T .

Because we have introduced a new set of goods, we must modify the household well-being functions by enlarging their domains. The obvious nota-

tion to express this extra dependence is to write $W_h(X_h, G, I, T_h)$ for the well-being of household h . Moreover, h 's feasible set F_h now also depends on G , I , and T_h ; so we write the set of feasible household choices as $F_h(G, I, T_h, X_{-h})$.

To try to determine the optimum public policy, imagine a two-stage game. The government has the first move, choosing T and then G and I from K_T . Households go second, reacting to decisions made by the government. Imagine that a social equilibrium $X^* = (X_1^*, \dots, X_H^*)$ is reached and that the equilibrium is unique. (We assume that if there are multiple equilibria, the government can select among them by resorting to public signals.) Clearly, this equilibrium X^* is a function of G , I , and T . An intelligent and benevolent government will anticipate it and choose T , G , and I from K_T in such a way as to maximize the resulting social well-being $V(W(X_1^*), \dots, W(X_H^*))$.

The public policy problem we have just designed, involving as it does a double optimization, is technically very difficult. It transpires, for example, that even in some of the simplest model economies one can imagine, $F_h(G, I, T_h, X_{-h})$ is not convex. This means that the social equilibrium cannot be guaranteed to depend continuously on G , I , and T , as was shown by Mirrlees in 1984. This in turn means that standard techniques are not suitable for the government's optimization problem. In fact, of course, even "double optimization" is a huge simplification. The government chooses; people respond by trading, producing, consuming; the government chooses again; people respond once again—and so forth in an unending series of moves and counter-moves. Identifying the optimum public policy involves severe computational difficulties.

6 Matters of Trust: Laws and Norms

The previous examples demonstrate that a fundamental problem facing people who would like to transact with one another concerns *trust*. For example, the extent to which parties trust one another shapes the sets F_h and K_T . If the parties do not trust one another, what could have been mutually beneficial transactions will not take place. But what grounds does a person have for trusting someone to do what he promises to do under the terms of an agreement? Such grounds can exist if promises can be made credible. Societies everywhere have constructed mechanisms to create credibility of this kind, but in different ways. What the mechanisms have in common, however, is that individuals

who fail to comply with agreements without a good reason are punished.

How does that common feature work?

In Becky's world the rules governing transactions are embodied in the law. The markets Becky's family enters are supported by an elaborate legal structure (a public good). Becky's father's firm, for example, is a legal entity; as are the financial institutions he deals with in order to accumulate his retirement pension, to save for Becky's and Sam's education, and so on. Even when someone in the family goes to the grocery store, the purchases (paid for with cash or by card) involve the law, which provides protection for both parties (the grocer, in case the cash is counterfeit or the card is void; the purchaser, in case the product turns out on inspection to be substandard). The law is enforced by the coercive power of the state. Transactions involve legal contracts backed by an *external enforcer*, namely, the state. It is because Becky's family and the grocery store's owner are confident that the government has the ability and willingness to enforce contracts (i.e., to continue to supply the public good in question) that they are willing to make transactions.

What is the basis of that confidence? After all, the contemporary world has shown that there are states and there are states. Why should Becky's family trust the government to carry out its tasks in an honest manner? A possible answer is that the government in her country worries about its *reputation*: a free and inquisitive press in a democracy helps to sober the government into believing that incompetence or malfeasance would mean an end to its rule come the next election. Notice how the argument involves a system of interlocking beliefs about the abilities and intentions of others. The millions of households in Becky's country trust their government (more or less!) to enforce contracts, because they know that government leaders know that not to enforce contracts efficiently would mean being thrown out of office. In their turn, each party to a contract trusts the other to refrain from reneging (again, more or less!), because each knows that the other knows that the government can be trusted to enforce contracts. And so on. Trust is maintained by the threat of punishment (a fine, a jail term, dismissal, or whatever) for anyone who breaks a contract. Once again, we are in the realm of equilibrium beliefs, held together by their own bootstraps. Mutual trust encourages people to seek out mutually beneficial transactions and engage in them. As the formal argument that supports the above claim is very similar to the one showing that

social norms contain mechanisms for enforcing agreements, we turn to the place of social norms in people's lives.

Although the law of contracts exists also in Desta's country, her family cannot depend on it because the nearest courts are far from their village. Moreover, there are no lawyers in sight. As transport is enormously costly, economic life is shaped outside a formal legal system. In short, crucial public goods and infrastructure are either unavailable, or, at best, in short supply. But even though there is no external enforcer, Desta's parents do make transactions with others. Credit (not dissimilar to insurance in her village) involves saying, "I will lend to you now if you promise to repay me when you can." Saving for funerals involves saying, "I agree to abide by the terms and conditions of the *iddir*." And so on. But why should the parties have any confidence that the agreements will not be broken?

Such confidence can be justified if agreements are *mutually enforced*. The basic idea is this: a credible threat by members of a community that stiff sanctions will be imposed on anyone who breaks an agreement can deter everyone from breaking it. The problem is then to make the threat credible. In Desta's world credibility is achieved by recourse to social norms of behavior.

By a *social norm* we mean a rule of behavior followed by members of a community. A rule of behavior (or "strategy" in economic parlance) reads like, "I will do X if you do Y," "I will do P if Q happens," and so forth. For a rule of behavior to *be* a social norm, it must be in the interest of everyone to act in accordance with the rule if all others act in accordance with it. Social norms are equilibrium rules of behavior. We will now see how social norms work and how transactions based on them compare with market-based transactions. To do this we will study insurance as a commodity.

7 Insurance

To insure oneself against a risk is to act in ways to reduce that risk. (Formally, a RANDOM VARIABLE [III.73 §4] \tilde{X} is said to be riskier than a random variable \tilde{Y} if there is a random variable \tilde{Z} with zero mean such that \tilde{X} has the same distribution as $\tilde{Y} + \tilde{Z}$. In this case, \tilde{X} and \tilde{Y} have the same mean but \tilde{X} is more "spread out.") As long as it does not cost too much, risk-averse households will want to reduce risk by purchasing insurance: in fact, avoiding risk would seem to be a universal urge. To formalize these notions, consider an

isolated village, such as Desta's. Suppose for simplicity that it contains H identical households. If household h 's food consumption is X_h (represented by a single real number), let us say that its well-being is $W(X_h)$. We shall assume that $W'(X_h) > 0$ (that is, more food leads to greater well-being) and that $W''(X_h) < 0$ (the more food you already have, the less you benefit from yet more). We shall confirm below that the second property of W , its strict concavity, implies, and is implied by, risk aversion; but the basic reason is simple: if W is strictly concave, then you gain less when you are lucky than you lose when you are unlucky.

For simplicity, let us suppose that the production of food by a household h , which is subject to chance factors such as the weather, involves no effort. Since the output is uncertain, we represent it by a random variable \tilde{X}_h , with expected value μ , which is assumed to be positive. We shall denote expectations by \mathbb{E} .

If a household h is completely self-sufficient, then its expected well-being is simply $\mathbb{E}(W(\tilde{X}_h))$. However, the strict concavity of W implies that $W(\mu) > \mathbb{E}(W(\tilde{X}_h))$. To put this in words: h 's well-being at the average level of production is greater than the expectation of h 's well-being if the production is random. This means that h will prefer a sure level of consumption to a risky one with mean equal to that sure level. In short, h is risk averse. Define a number $\bar{\mu}$ by $W(\bar{\mu}) = \mathbb{E}(W(\tilde{X}_h))$. So $\bar{\mu}$ is the level of production that achieves the expected well-being. This will be less than μ , and so $\mu - \bar{\mu}$ is a measure of the cost of the risk that a self-sufficient household bears. Notice that the greater the "curvature" of W is, the greater the cost is of the risk associated with \tilde{X}_h . (A useful measure of curvature turns out to be $-XW''(X)/W'(X)$. We will make use of this measure when discussing intertemporal choices.) To see how households could gain by pooling their risks, let us write $\tilde{X}_h = \mu + \tilde{\varepsilon}_h$, where $\tilde{\varepsilon}_h$ is a random variable with mean zero, variance σ^2 , and finite support. Suppose for simplicity that the random variables $\tilde{\varepsilon}_h$ are identical (i.e., they do not depend on h). Let the correlation coefficient of any two of these distributions be ρ . It turns out that, as long as $\rho < 1$, households can reduce their risks by agreeing to share their outputs. Suppose that households are able to observe one another's outputs. Given that the random variables \tilde{X}_h are identical, the obvious insurance scheme is to share out the outputs equally. Under this scheme, h 's uncertain food consumption becomes the average of $\tilde{X}_1, \dots, \tilde{X}_H$, which is an improvement on self-sufficiency because $\mathbb{E}(W(\sum \tilde{X}_h / H)) > \mathbb{E}(W(\tilde{X}_h))$. The problem is

that, without an enforcement mechanism, the agreement to share will not stick, because once each household knows how much food every household has produced, all but the unluckiest households will wish to renege. To see why, notice first that the luckiest households will renege because their outputs are above the average; but this means that the next luckiest set of households will renege because their outputs are above the reduced average; and so on, down to the unluckiest households. Since households know in advance that this will happen if there are no enforcement mechanisms, they will not enter the scheme in the first place: the only social equilibrium is pure self-sufficiency and there is no pooling of risk.

Let us call the insurance game just described the *stage game*. Although pure self-sufficiency is the only social equilibrium for the stage game, we shall now see that the situation changes if the game is played repeatedly. To model this, let us use the letter t to denote time, and let us take time to be a nonnegative integer. (The game might, for instance, take place every year, with 0 standing for the current year.) Let us assume that the villagers face the same set of risks in each time period, and that the risk in one year is independent of the risks in all other years. Also assume that, in each period, once food outputs are realized, households decide independently of one another whether they will abide by the agreement to share their produce equally or whether they will renege on it.

Although future well-being is important to a household, it will typically be less important than present well-being. To model this we introduce a positive parameter δ , which measures how much a household discounts its future well-being. The assumption is that, when making calculations at $t = 0$, a household divides its well-being at time t by a factor $(1 + \delta)^t$: that is, the importance decays by a certain fixed percentage at each time period. We shall now show that, provided δ is sufficiently small (i.e., provided that households care enough about their future well-being), there is a social equilibrium in which households abide by the agreement to share their aggregate output equally.

Let $\tilde{Y}_h(t)$ be the uncertain amount of food available to household h at time t . If all households are participating in the agreement, then $\tilde{Y}_h(t)$ will be $\mu + (\sum \tilde{\varepsilon}_{h'})/H$, and if there is no agreement, then it will be $\mu + \tilde{\varepsilon}_h$. At time $t = 0$ the total expected well-being of household h , present and future, is $\sum_0^\infty \mathbb{E}(W(\tilde{Y}_h(t)))/(1 + \delta)^t$. (To calculate this we took, for each $t \geq 0$, the expected well-being of h at time t

and divided it by $(1 + \delta)^t$. Then we added these numbers up.)

Now consider the following strategy that h might adopt: it begins by participating in the insurance scheme and continues to participate so long as no household has reneged on the agreement; but it withdraws from the scheme from the date following the first violation of the agreement by some household. Game theorists have christened this the “grim strategy,” or simply *grim*, because of its unforgiving nature. Let us see how grim could support the original agreement to share aggregate output equally at every date. (For a general account of repeated games and the variety of social norms that can sustain agreements, see Fudenberg and Maskin (1986).)

Suppose that household h believes that all other households have chosen grim. Then h knows that none of the other households will be the first to defect. What should h do then? We will show that if δ is small enough, h can do no better than play grim. As the same reasoning would be applicable to all other households, we should conclude that, for small enough values of δ , grim is an equilibrium strategy in the repeated game. But if all households play grim, then no household will ever defect. Grim can therefore function as a social norm for sustaining cooperation. Let us see how the argument works.

The basic idea is simple. As all other households are assumed to be playing grim, household h would enjoy a one-period gain by defecting if its own output exceeded the average output of all households. But if h defects in any period, all other households will defect in all following periods (they are assumed to be playing grim, remember). Therefore, h 's own best option in all following periods will be to defect also, which means that subsequent to a single deviation by h , the outcome can be predicted to be pure self-sufficiency. So, set against a one-period gain that household h would enjoy if its output exceeded the average output of all households is the loss it would suffer from the following date because of the breakdown of cooperation. That loss exceeds the one-period gain if δ is small enough. So, if δ is sufficiently small, household h will not defect, but will adopt grim; implying that grim is an equilibrium strategy and equal sharing among households in every period is a social equilibrium.

To formalize the above argument, we consider the situation in which h 's incentive to defect is greatest. Let A and B be the minimum and maximum possible outputs

of any household. Then the maximum gain that household h could possibly enjoy from defecting at $t = 0$ arises if h happens to produce B and all other households happen to produce A . Since the average output in this eventuality is $(B + (H - 1)A)/H$, the one-period gain that household h would enjoy from defecting is

$$W(B) - W\left(\frac{B + (H - 1)A}{H}\right).$$

But h knows that if it defects, the expected loss in each subsequent period (i.e., from $t = 1$ onward) will be $\mathbb{E}(W(\sum \tilde{X}_{h'}/H)) - \mathbb{E}(W(\tilde{X}_h))$. In order to simplify the notation, let us write $\mathbb{E}(W(\sum \tilde{X}_{h'}/H)) - \mathbb{E}(W(\tilde{X}_h))$ as L . Household h can then calculate that the expected total loss it will suffer from defecting at $t = 0$ is $L \sum_{t=1}^{\infty} (1 + \delta)^{-t}$, which equals L/δ . If this future loss exceeds the present gain from defecting, then household h will not want to defect. In other words, h will not want to defect if

$$\frac{L}{\delta} > W(B) - W\left(\frac{B + (H - 1)A}{H}\right)$$

or

$$\delta < L / \left(W(B) - W\left(\frac{B + (H - 1)A}{H}\right) \right). \quad (1)$$

But if h does not find it in its interest to defect when the one-period gain from defection is the largest possible, it will certainly not want to defect in any other situation. We conclude that if inequality (1) holds, then grim is an equilibrium strategy and equal sharing among households in every period is a resulting social equilibrium. Notice that, as we said, this will happen if δ is sufficiently small.

We usually reserve the term “society” to denote a collective that has managed to find a mutually beneficial equilibrium. Notice, however, that another social equilibrium of the repeated game is each household for itself. If everyone believed that all others would break the agreement from the start, then everyone would break the agreement from the start. Noncooperation would involve each household selecting the strategy: renege on the agreement. Failure to cooperate could be due simply to a collection of unfortunate, self-confirming beliefs, and nothing else. It is also easy to show that noncooperation is the only social equilibrium of the repeated game if

$$\delta > L / \left(W(B) - W\left(\frac{B + (H - 1)A}{H}\right) \right). \quad (2)$$

We now have in hand a tool for understanding how a community can slide from cooperative to noncooperative behavior. For example, political instability (in the extreme, civil war) can mean that households are

increasingly concerned that they will be forced to disperse from their village. This translates into an increase in δ . Similarly, if households fear that their government is now bent on destroying communal institutions in order to strengthen its own authority, δ will increase. But from (1) and (2) we know that if δ increases sufficiently, then cooperation ceases. The model therefore offers an explanation for why, in recent decades, cooperation at the local level has declined in the unsettled regions of sub-Saharan Africa. Social norms work only when people have reasons to value the future benefits of cooperation.

In the above analysis, we allowed for the possibility that, in each period, household risks were positively correlated. Moreover, the number of households in any village is typically not large. These are two reasons why Desta's household is unable to attain anything like full insurance against the risk they face. Becky's parents, in contrast, have access to an elaborate set of insurance markets that pool the risks of hundreds of thousands of households across the country (even the world, if the insurance company is a multinational). This helps to reduce individual risk more than Desta's parents can, because, first, spatially distant risks are more likely to be uncorrelated, and, second, Becky's parents can pool their risk with many more households. With enough households and enough independence of their risk, THE LAW OF LARGE NUMBERS [III.73 §4] practically guarantees that equal sharing among those households will provide each one with the average μ . This is an advantage of markets, backed by the coercive power of the state as an external enforcer: in a competitive market, insurance contracts are available, enabling people who do not know one another to do business through third parties, in this case the insurance companies.

Many of the risks that Desta's parents face, such as low rainfall, will in fact be very similar for all households in their village. Since the insurance they are able to obtain within their village is therefore very limited, they adopt additional risk-reducing strategies, such as diversifying their crops. Desta's parents plant maize, *tef*, and *enset* (an inferior crop), with the hope that even if maize were to fail one year, *enset* would not let them down. That the local resource base in Desta's village is communally owned probably also has something to do with a mutual desire to pool risks. Woodlands are spatially nonhomogeneous ecosystems. In one year one group of plants bears fruit, in another year some other group does. If the woodland were divided into private parcels, each household would face a greater risk than

it would under communal ownership. The reduction in individual household risks owing to communal ownership may be small, but as average incomes are very low, household benefits from communal ownership are large. (For a fuller account of the management of local commons in poor countries, see Dasgupta (1993).)

8 The Reach of Transactions and the Division of Labor

Payments in Becky's world are made in money, expressed in U.S. dollars. Money would not be required in a world where everyone was known to be utterly trustworthy, people did not incur computational costs, and transactions were costless: simple IOUs, stipulating repayment in terms of specific good and services, would suffice in that world. However, we do not live in that world. A debt in Becky's world involves a contract specifying that the borrower is to receive a certain number of dollars and that he promises to repay the lender dollars in accordance with an agreed schedule. When signing the contract the relevant parties entertain certain beliefs about the dollar's future value in terms of goods and services. Those beliefs are in part based on their confidence in the U.S. government to manage the value of the dollar. Of course, the beliefs are based on many other things as well; but the important point remains that money's value is maintained only because people believe it will be maintained (the classic reference on this is Samuelson (1958)). Similarly, if, for whatever reason, people feared that the value would not be maintained, then it would not be maintained. Currency crashes, such as the one that occurred in Weimar Germany in 1922–23, are an illustration of how a loss in confidence can be self-fulfilling. Bank runs share that feature, as do stock market bubbles and crashes. To put it formally, there are multiple social equilibria, each supported by a set of self-fulfilling beliefs.

The use of money enables transactions to be anonymous. Becky frequently does not know the salespeople in the department stores of her town's shopping mall, nor do they know Becky. When Becky's parents borrow from their bank, the funds made available to them come from unknown depositors. Literally millions of transactions occur each day between people who have never met and will never meet in the future. The problem of creating trust is solved in Becky's world by building confidence in the medium of exchange: money. The value of money is maintained by the state, which has an incentive to maintain it because, as we saw earlier, it

wishes not to destroy its reputation and be thrown out of office.

In the absence of infrastructure, markets are unable to penetrate Desta's village. Becky's suburban town, by contrast, is embedded in a gigantic world economy. Becky's father is able to specialize as a lawyer only because he is assured that his income can be used to purchase food in the supermarket, water from the tap, and heat from cooking ovens and radiators. Specialization enables people to produce more in total than they would be able to if they were each required to diversify their activities. Adam Smith famously remarked that the division of labor is limited by the size of the market. Earlier we noted that Desta's household does not specialize, but produces pretty much all of its daily requirements from a raw state. Moreover, the many transactions it enters into with others, being supported by social norms, are of necessity personalized, thus limited. There is a world of a difference between laws and social norms as the basis of economic activities.

9 Borrowing, Saving, and Reproducing

If you do not have insurance, then your consumption will depend heavily on various contingencies. Purchasing insurance helps to smooth out this dependence. We shall see presently that the human desire to smooth out the dependence on contingencies is related to the equally common desire to smooth out consumption across time: they are both a reflection of the strict concavity of the well-being function W . The flow of income over a person's lifetime tends not to be smooth, so people look for mechanisms, such as mortgages and pensions, that enable them to transfer consumption across time. For instance, Becky's parents took out a mortgage on their house because at the time of purchase they did not have sufficient funds to finance it. The resulting debt decreased their future consumption, but it enabled them to buy the house at the time they did and thereby raise current consumption. Becky's parents also pay into a pension fund, which transfers present consumption to their retired future. Borrowing for current consumption transfers future consumption to the present; saving achieves the reverse. Since capital assets are productive, they can earn positive returns if they are put to good use. This is one reason why, in Becky's world, borrowing involves having to pay interest, while saving and investing earn positive returns.

Becky's parents also make a considerable investment in their children's education, but they do not expect to

be repaid for this. In Becky's world, resources are transferred from parents to children. Children are a direct source of parental well-being; they are not regarded as investment goods.

A simple way to formulate the problem Becky's parents face when they arrange transfers of resources across time is to imagine that they view themselves as part of a dynasty. This means that, in reaching their consumption and saving decisions, they take explicit note not only of their own well-being and the well-being of Becky and Sam, but also of the well-being of their potential grandchildren, great grandchildren, and so on, down the generations.

To analyze the problem, it is notationally tidier to assume that time is a continuous variable. At time t (which we take to be greater than or equal to 0), let $K(t)$ denote household wealth and $X(t)$ the consumption rate, which is some aggregate based on the market prices of what they consume. In practice, a household will want to smooth its consumption across both time and contingencies, but in order to concentrate on time we shall consider a deterministic model. Suppose that the market rate of return on investment is a positive constant r . This means that if household wealth at time t is $K(t)$, then the income it earns from that wealth at t is $rK(t)$. The dynamical equation describing the dynasty's consumption options over time is then

$$dK(t)/dt = rK(t) - X(t). \quad (3)$$

The right-hand side of the equation is the difference between the dynasty's investment income at time t (which is r times its wealth at t) and its consumption at t . This amount is saved and invested, so it gives the rate of increase of the dynasty's wealth at t . The present time is $t = 0$ and $K(0)$ is the wealth that Becky's parents have inherited from the past. Earlier, we assumed that the household allocates its consumption across contingencies by maximizing its expected well-being. The corresponding quantity for allocating consumption across time is

$$\int_0^\infty W(X(t))e^{-\delta t} dt, \quad (4)$$

where, as before, we assume that W satisfies the conditions $W'(X) > 0$ and $W''(X) < 0$. The parameter δ is once again a measure of the rate at which future well-being is discounted—owing to shortsightedness, the possibility of dynastic extinction, and so on. The difference between this and the previous δ is that now we are considering a continuous model rather than a

discrete one, but the decay is still assumed to be exponential. In Becky's world the rate of return on investment is large; that is, investment is very productive. So it makes empirical sense to suppose that $r > \delta$. We will see presently that this condition provides Becky's parents with the incentive to accumulate wealth and pass it on to Becky and Sam, who in turn will accumulate their wealth and pass *that* on, and so on. For simplicity, let us suppose that the "curvature" of W , which is $-XW''(X)/W'(X)$, is equal to a parameter α , whose value exceeds 1.³ As we saw earlier, strict concavity of W means that you gain less from increasing consumption than you lose from decreasing it by the same amount. The strength of this effect is measured by α : the larger it is, the greater the benefit of any smoothing you are able to do.

Becky's parents' problem at $t = 0$ is to maximize the quantity in (4) by making a suitable choice of the rate at which they consume their wealth (namely, $X(t)$), subject to the condition (3), together with the conditions that $K(t)$ and $X(t)$ should not be negative.⁴ This is a problem in the CALCULUS OF VARIATIONS [III.96]. But it is of a somewhat unusual form, in that the horizon is infinite and there is no boundary condition at infinity. The reason for the latter is that Becky's parents would ideally like to *determine* the level of assets that the dynasty ought to aim at in the long run; they do not think it is appropriate to specify it in advance. If we assume for the moment that a solution to the optimization problem exists, then it turns out that it must satisfy the *Euler-Lagrange equation*:

$$\alpha(dX(t)/dt) = (r - \delta)X(t), \quad t \geq 0. \quad (5)$$

This equation is easily solved, and gives

$$X(t) = X(0)e^{(r-\delta)t/\alpha}. \quad (6)$$

3. This means that W has the form $B - AX^{-(\alpha-1)}$, where A (which is a positive number) and B (which can be of either sign) are the two arbitrary constants that arise when we integrate the curvature of W to arrive at W itself. We will see presently that the values that are adopted for A and B have no bearing on the decisions that Becky's parents will want to make; that is, Becky's parents' optimum decision is independent of A and B . Notice that, as $\alpha > 1$, $W(X)$ is bounded above. The above form is particularly useful in applied work, because in order to estimate $W(X)$ from data on household consumption, one has to estimate only one parameter, α . Empirical studies of saving behavior in the United States have revealed that α is in the range 2–4.

4. This problem originated in a classic paper by Ramsey (1928). Ramsey insisted that $\delta = 0$ and devised an ingenious argument to show that an optimum function $X(t)$ exists despite the fact that the integral in (4) does not converge. For simplicity, I am assuming $\delta > 0$. As $W(X)$ is bounded above and $r > 0$ (meaning that it is feasible for $X(t)$ to grow indefinitely), we should expect (4) to converge if $X(t)$ is allowed to rise fast enough.

However, we are free to choose $X(0)$. Koopmans showed in 1965 that $X(t)$ in (6) is optimal if $W'(X(t))K(t)e^{-\delta t} \rightarrow 0$ as $t \rightarrow \infty$. It transpires that, for the model in hand, there is a value of $X(0)$, which we shall write as $X^*(0)$, such that the condition (3) and Koopmans's asymptotic condition are satisfied by the function $X(t)$ given in (6). This implies that $X^*(0)e^{(r-\delta)t/\alpha}$ is the unique optimum. Consumption grows at the percentage rate $(r - \delta)/\alpha$ and dynastic wealth accumulates continually in order to make that rising consumption level possible. All other things being equal, the larger the productivity of investment r , the higher the optimum rate of growth of consumption. By contrast, the larger the value of α , the lower the rate of growth of consumption, since there is a greater wish to spread it out among the generations.

Let us conduct a simple exercise with our finding. Suppose the annual market rate of return is 4% (i.e., $r = 0.04$ per year)—a reasonable figure for the United States—that δ is small, and that $\alpha = 2$. Then we can conclude from (6) that optimum consumption will grow at an annual rate of 2%; meaning that it will double every thirty-five years—roughly, every generation. The figure is close to the postwar growth experience in the United States.

For Desta's parents the calculations are very different, since they are heavily constrained in their ability to transfer consumption across time. For example, they have no access to capital markets from which they can earn a positive return. Admittedly, they invest in their land (clearing weeds, leaving portions fallow, and so forth), but that is to prevent the productivity of the land from declining. Moreover, the only way they are able to draw on the maize crop following each harvest is to store it. Let us see how Desta's household would ideally wish to consume that harvest over the annual cycle.

Let $K(0)$ be the harvest, measured, say, in kilocalories. As rats and moisture are a potent combination, stocks depreciate. If $X(t)$ is the planned rate of consumption and γ the rate of depreciation of the maize stock, then the stock at t satisfies the equation

$$dK(t)/dt = -X(t) - \gamma K(t). \quad (7)$$

Here, γ is assumed to be positive and both $X(t)$ and $K(t)$ nonnegative. Imagine that Desta's parents regard their household's well-being over the year to be $\int_0^1 W(X(t)) dt$. As with Becky's household, let $-XW''(X)/W'(X)$ be equal to a number $\alpha > 1$. Desta's parents' optimization exercise is to maximize

$\int_0^1 W(X(t)) dt$, subject to (7) and the condition that $K(1) \geq 0$.

This is a straightforward problem in the calculus of variations. It can be shown that the optimum maize consumption *declines* over time at the rate γ/α . This explains why Desta's family consume less and become physically weaker as the next harvest grows nearer. But Desta's parents have realized that the human body is a more productive bank. So the family consumes a good deal of maize during the months following each harvest so as to accumulate body mass, but they draw on that reserve during the weeks before the next harvest, when maize reserves have been depleted. Across the years maize consumption assumes a sawtooth pattern. (Readers may wish to construct the model that incorporates the body as a store of energy: see Dasgupta (1993) for details.)

As Desta and her siblings contribute to daily household production, they are economically valuable assets. Her male siblings, however, offer a higher return to their parents, because the custom (itself a social equilibrium!) is for girls to leave home on marriage and for boys to inherit the family property and offer security to their parents in old age. Because of an absence of capital markets and state pensions, male children are an essential form of investment. The transfer of resources in Desta's household, in contrast to Becky's, will be from the children to their parents.

The under-five mortality rate in Ethiopia was, until relatively recently, in excess of 300 per 1000 births. So, parents had to aim at large families if they were to have a reasonable chance of being looked after by a male child in their old age. But fertility is not entirely a private matter, since people are influenced by the choices of others. This gives rise to a certain inertia in household behavior even under changing circumstances, which is why even though the under-five mortality rate has fallen in Ethiopia in recent decades, Becky has five siblings.⁵ High population growth has placed additional pressure on the local ecosystem, meaning that the local commons that used to be managed in

a sustainable manner no longer are. That they are not is reflected in Desta's mother's complaint that the daily time and effort required to collect from the local commons has increased in recent years.

10 Differences in Economic Life among Similar People

In this article, I have used Becky's and Desta's experiences to show how it can be that the lives of essentially very similar people can become so different (for further elaboration, see Dasgupta (2004)). Desta's life is one of poverty. In her world people do not enjoy food security, do not own many assets, are stunted and wasted, do not live long (life expectancy at birth in Ethiopia is under fifty years), cannot read or write, are not empowered, cannot insure themselves well against crop failure or household calamity, do not have control over their own lives, and live in unhealthy surroundings. The deprivations reinforce one another, so that the productivity of labor effort, ideas, physical capital, and of land and natural resources are all very low and remain low. The rate of return on investment is zero, perhaps even negative (as it is with the storage of maize). Desta's life is filled with *problems* each day.

Becky suffers from no such deprivation (for example, life expectancy at birth in the United States is nearly eighty years). She faces what her society calls *challenges*. In her world, the productivity of labor effort, ideas, physical capital, and of land and natural resources are all very high and continually increasing; success in meeting each challenge reinforces the prospects of success in meeting further challenges.

We have seen, however, that, despite the enormous differences between Becky's and Desta's lives, there is a unified way to view them, and that mathematics is an essential language for analyzing them. It is tempting to pronounce that life's essentials cannot be reduced to mere mathematics; but in fact mathematics is essential to economic reasoning. It is essential because in economics we deal with quantifiable objects of vital interest to people.

Acknowledgments. In describing Desta's life, I have received much guidance from my colleague Pramila Krishnan.

Further Reading

Dasgupta, P. 1993. *An Inquiry into Well-Being and Destitution*. Oxford: Clarendon Press.

5. See Dasgupta (1993) for the use of interdependent preferences to explain fertility behavior. In the notation of the section on social equilibria, we are to suppose that household h 's well-being has the form $W_h(X_h, X_{-h})$, where one of the components of X_h is the number of births in the household, and that the higher the fertility rate is among other households in the village, the larger the desired number of children in h . The theory based on interdependent preferences interprets transitions from high to low fertility rates as bifurcations. Fertility rates are expected to decline even in Ethiopia. Interdependent preferences are currently being much studied by economists (see Durlauf and Young 2001).

- . 2004. World poverty: causes and pathways. In *Annual World Bank Conference on Development Economics 2003: Accelerating Development*, edited by F. Bourguignon and B. Pleskovic, pp. 159–96. New York: World Bank and Oxford University Press.
- Debreu, G. 1959. *Theory of Value*. New York: John Wiley.
- Diamond, J. 1997. *Guns, Germs and Steel: A Short History of Everybody for the Last 13,000 Years*. London: Chatto & Windus.
- Durlauf, S. N., and H. Peyton Young, eds. 2001. *Social Dynamics*. Cambridge, MA: MIT Press.
- Evans, G., and S. Honkapohja. 2001. *Learning and Expectations in Macroeconomics*. Princeton, NJ: Princeton University Press.
- Fogel, R. W. 2004. *The Escape from Hunger and Premature Death, 1700–2100: Europe, America, and the Third World*. Cambridge: Cambridge University Press.
- Fudenberg, D., and E. Maskin. 1986. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* 54(3):533–54.
- Landes, D. 1998. *The Wealth and Poverty of Nations*. New York: W. W. Norton.
- Ramsey, F. P. 1928. A mathematical theory of saving. *Economic Journal* 38:543–49.
- Samuelson, P. A. 1947. *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.
- . 1958. An exact consumption loan model with or without the social contrivance of money. *Journal of Political Economy* 66:1002–11.

VII.9 The Mathematics of Money

Mark Joshi

1 Introduction

The last twenty years have seen an explosive growth in the use of mathematics in finance. Mathematics has made its way into finance mainly via the application of two principles from economics: *market efficiency* and *no arbitrage*.

Market efficiency is the idea that the financial markets price every asset correctly. There is no sense in which a share can be a “good buy,” because the market has already taken all available information into account. Instead, the only way that we have of distinguishing between two assets is their differing *risk characteristics*. For example, a technology share might offer a high rate of growth but also a high probability of losing a lot of money, while a U.K. or U.S. government bond would offer a much smaller rate of growth, but an extremely low probability of losing money. In fact, the probability of loss is so small in the latter case

that these instruments are generally regarded as being riskless.

No arbitrage, the second fundamental principle, simply says that it is impossible to make money without taking risk. It is sometimes called the “no free lunch” principle. In this context, “making money” is defined to mean making *more* money than could be obtained by investing in a riskless government bond. A simple application of the principle of no arbitrage is that if one changes dollars into yen and then the yen into euros and then the euros back into dollars, then, apart from any transaction costs, one will finish with the same number of dollars that one started with. This forces a simple relationship between the three foreign exchange (FX) rates:

$$FX_{\$,€} = FX_{\$,¥}FX_{¥,€}. \quad (1)$$

Of course, occasional anomalies and exceptions to this relationship can occur, but these will be spotted by traders. The exploitation of the resulting arbitrage opportunity will quickly move the exchange rates until the opportunity disappears.

One can roughly divide the use of mathematics in finance into four main areas.

Derivatives pricing. This is the use of mathematics to price *securities* (i.e., financial instruments), whose value depends purely upon the behavior of another asset. The simplest example of such a security is a *call option*, which is the right, but not the obligation, to buy a share for a pre-agreed price, K , on some specified future date. The pre-agreed price is called the *strike*. The pricing of derivatives is heavily reliant upon the principle of no arbitrage.

Risk analysis and reduction. Any financial institution has holdings and borrowings of assets; it needs to keep careful control of how much money it can lose from adverse market moves and to reduce these risks as necessary to keep within the owners’ desired risk profiles.

Portfolio optimization. Any investor in the markets will have notions of how much risk he wants to take and how much return he wants to generate, and most importantly of where he sees the trade-off between the two. There is, therefore, a theory of how to invest in shares in such a way as to maximize the return at a given level of risk. This theory relies greatly on the principle of market efficiency.

Statistical arbitrage. Crudely put, this is using mathematics to predict price movements in the stock market, or indeed in any other market. Statistical arbitrage

T&T note: check word spacing before press.

trageurs laugh at the concept of market efficiency, and their objective is to exploit the inefficiencies in the market to make money.

Of these four areas, it is derivatives pricing that has seen the greatest growth in recent years, and which has seen the most powerful application of advanced mathematics.

2 Derivatives Pricing

2.1 Black and Scholes

Many of the foundations of mathematical finance were laid down by Bachelier (1900) in his thesis; his mathematical study of BROWNIAN MOTION [IV.24] preceded that of Einstein (see Einstein (1985), which contains his 1905 paper). However, his work was neglected for many years and the great breakthrough in derivatives pricing was made by Black and Scholes (1973). They showed that, under certain reasonable assumptions, it was possible to use the principle of no arbitrage to guarantee a unique price for a call option. The pricing of derivatives had ceased to be an economics problem and had become a mathematics problem.

The result of Black and Scholes was deduced by extending the principle of no arbitrage to encompass the idea that an arbitrage could result not just from static holdings of securities, but also from continuously trading them in a dynamic fashion depending upon their price movements. It is this principle of no *dynamic* arbitrage that underpins derivatives pricing.

In order to properly formulate the principle, we have to use the language of probability theory.

An *arbitrage* is a trading strategy in a collection of assets, the *portfolio*, such that

- (i) initially the portfolio has a value of zero;
- (ii) the probability that the portfolio will have a negative value in the future is zero;
- (iii) the probability that the portfolio will have a positive value in the future is greater than zero.

Note that we do not require the profit to be certain; we merely require that it is possible that money may be made with no risk taken. (Recall that the notion of making money is by comparison with a government bond. The same is true of the “value” of a portfolio: it will be considered positive in the future if its price has increased by more than that of a government bond.)

The prices of shares appear to fluctuate randomly, but often with a general upward or downward tendency. It is natural to model them by means of a Brownian motion with an extra “drift term.” This is what Black and Scholes did, except that it was the *logarithm* of the share price $S = S_t$ that was assumed to follow a Brownian motion W_t with a drift. This is a natural assumption to make, because changes in prices behave multiplicatively rather than additively. (For example, we measure inflation in terms of percentage increases.) They also assumed the existence of a riskless bond, B_t , growing at a constant rate. To put these assumptions more formally:

$$\log S = \log S_0 + \mu t + \sigma W_t, \quad (2)$$

$$B_t = B_0 e^{rt}. \quad (3)$$

Notice that the expectation of $\log S$ is $\log S_0 + \mu t$, so it changes at a rate μ , which is called the *drift*. The term σ is known as the *volatility*. The higher the volatility, the greater the influence of the Brownian motion W_t , and the more unpredictable the movements of S . (An investor will want a large μ and a small σ ; however, market efficiency ensures that such shares are rather rare.) Under additional assumptions such as that there are no transaction costs, that trading in a share does not affect its price, and that it is possible to trade continuously, Black and Scholes showed that if there is no dynamic arbitrage, then at time t , the price of a call option, $C(S, t)$, that expires at time T must be equal to

$$BS(S, t, r, \sigma, T) = S\Phi(d_1) - Ke^{-r(T-t)}\Phi(d_2), \quad (4)$$

with

$$d_1 = \frac{\log(S/K) + (r + \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}} \quad (5)$$

and

$$d_2 = \frac{\log(S/K) + (r - \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}}. \quad (6)$$

Here, $\Phi(x)$ denotes the probability that a standard normal random variable has value less than x . As x tends to ∞ , $\Phi(x)$ tends to 1, and as x tends to $-\infty$, $\Phi(x)$ tends to 0. If we let t tend to T , we find that d_1 and d_2 tend to ∞ if $S_T > K$ (in which case $\log(S_T/K) > 0$) and to $-\infty$ if $S_T < K$. It follows that the price $C(S, t)$ converges to $\max(S_T - K, 0)$, which is the value of a call option at expiry, just as one would expect. We illustrate this in figure 1.

There are a number of interesting aspects to this result that go far beyond the formula itself. The first and most important result is that the price is unique. Using just the hypothesis that it is impossible to make a

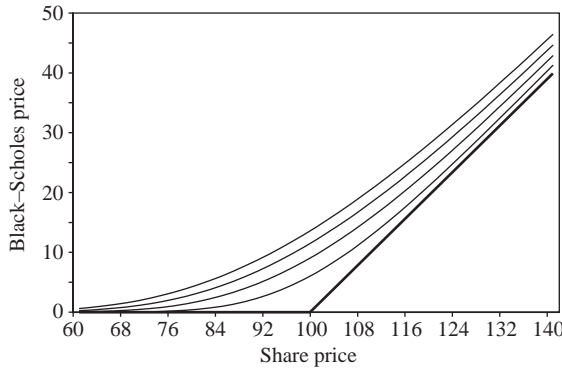


Figure 1 The Black-Scholes price of a call option struck at 100 for various maturities. The value decreases as maturity decreases, with the bottom line denoting a maturity of zero.

riskless profit, along with some natural and innocuous assumptions, we discover that there is only one possible price for the option. This is a very strong conclusion. It is not just the case that the option is a bad deal if traded at a different price: if a call option is bought for less or sold for more than the Black-Scholes price, then a *riskless profit* can be made.

A second fact, which may seem rather paradoxical, is that μ , the drift, does not appear anywhere in the Black-Scholes formula. This means that the expected behavior of the share's future mean price does not affect the price of the call option; our beliefs about the probability that the option will be used do not affect its price. Instead, it is the volatility of the share price that is all-important.

As part of their proof, Black and Scholes showed that the call option price satisfied a certain partial differential equation (PDE) now known as the *Black-Scholes equation*, or BS equation for short:

$$\frac{\partial C}{\partial t} + rS \frac{\partial C}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} = rC. \quad (7)$$

This part of the proof did not rely on the derivative being a call option: there is in fact a large class of derivatives whose prices satisfy the BS equation, differing only in boundary conditions. If one changes variables, setting $\tau = T - t$ and $X = \log S$, then the BS equation becomes THE HEAT EQUATION [I.3 §5.4] with an extra first-order term which can easily be removed. This means that the value of an option behaves in a similar way to time-reversed heat: it diffuses and spreads out the farther back one gets from the option's expiry and the more uncertainty there is about the value of the share at time T .

2.2 Replication

The fundamental idea underlying the Black-Scholes proof and much of modern derivatives pricing is *dynamic replication*. Suppose we have a derivative Y that pays an amount that depends on the value of the share at some set of times $t_1 < t_2 < \dots < t_n$, and suppose that the payout occurs at a certain time $T \geq t_n$. This can be expressed in terms of a *payoff function*, $f(t_1, \dots, t_n)$.

The value of Y will vary with the share price. If, in addition, we hold just the right number of the shares themselves, then a portfolio consisting of Y and the shares will be instantaneously immune to changes in the share price, i.e., its value will have zero rate of change with respect to the share price. As the value of Y will vary with time and share price, we will need to continuously buy and sell shares to maintain this neutrality to share-price movements. If we have sold a call option, then it turns out that we will have to buy when the share price goes up and sell when it goes down; so these transactions will cost us a certain amount of money.

Black and Scholes's proof showed that this sum of money was always the same and that it could be computed. The sum of money is such that by investing it in shares and riskless bonds, one can end up with a portfolio precisely equal in value to the payoff of Y no matter what the share price did in between.

Thus if one could sell Y for more than this sum of money, one would simply carry out the trading strategy from their proof and always end up ahead. Similarly, if one can buy Y for less, one does the negative of the strategy and always ends up ahead. Both of these are outlawed by the principle of no arbitrage, and a unique price is guaranteed.

The property that the payoff of any derivative can be replicated is called *market completeness*.

2.3 Risk-Neutral Pricing

A curious aspect of the Black-Scholes result, mentioned above, is that the price of a derivative does not depend upon the drift of the share price. This leads to an alternative approach to derivatives pricing theory called *risk-neutral pricing*. An arbitrage can be thought of as the ultimate unfair game: the player can only make money. By contrast, a MARTINGALE [IV.24 §4] encapsulates the notion of a fair game: it is a random process whose expected future value is always equal to its current value. Clearly, an arbitrage portfolio can never be

a martingale. So if we can arrange for everything to be a martingale, there can be no arbitrages, and the price of derivatives must be free of arbitrage.

Unfortunately, this cannot be done because the price of the riskless bond grows at a constant rate, and is therefore certainly not a martingale. However, we can carry out the idea for *discounted prices*: that is, for prices of assets when they are divided by the price of the riskless bond.

In the real world, we do not expect discounted prices to be martingales. After all, why buy shares if their mean return is no better than that of a bond that carries no risk? Nevertheless, there is an ingenious way of introducing martingales into the analysis: by changing the PROBABILITY MEASURE [III.73 §2] that one uses.

If you look back at the definition of arbitrage, you will see that it depends only on which events have zero probability and which have nonzero probability. Thus, it uses the probability measure in a rather incomplete way. In particular, if we use a different probability measure for which the sets of measure zero are the same, then the set of arbitrage portfolios will not change. Two measures with the same sets of measure zero are said to be *equivalent*.

A theorem of Girsanov says that if you change the drift of a Brownian motion, then the measure that you derive from it will be equivalent to the measure you had before. This means that we can change the term μ . A good value to choose turns out to be $\mu = r - \frac{1}{2}\sigma^2$.

With this value of μ , one has

$$\mathbb{E}(S/B_t) = S/B_0 \quad (8)$$

for any t , and since we can take any time as our starting point, it follows that S/B_t is a martingale. (The extra $-\frac{1}{2}\sigma^2$ in the drift comes from the concavity of the coordinate change to log-space.) This means that the expectation has been taken in such a way that shares do not carry any greater return, on average, than bonds. Normally, as we have mentioned, one would expect an investor to demand a greater return from a risky share than from a bond. (An investor who does not demand such compensation is said to be *risk neutral*.) However, now that we are measuring expectations differently, we have managed to build an equivalent model in which this is no longer the case.

This yields a way of finding arbitrage-free prices. First, pick a measure in which the discounted price processes of all the fundamental instruments, e.g., shares and bonds, are martingales. Second, set the discounted price process of derivatives to be the expectations

of their payoff; this makes them into martingales by construction.

Everything is now a martingale and there can be no arbitrage. Of course, this merely shows that the price is nonarbitrageable, rather than that it is the *only* nonarbitrageable price. However, work by Harrison and Kreps (1979) and by Harrison and Pliska (1981) shows that if a system of prices is nonarbitrageable, then there must be an equivalent martingale measure. Thus the pricing problem is reduced to classifying the set of equivalent martingale measures. Market completeness corresponds to the pricing measure being unique.

Risk-neutral evaluation has become such a pervasive technique that it is now typical to start a pricing problem by postulating risk-neutral dynamics for assets rather than real-world ones.

We now have two techniques for pricing: the Black-Scholes replication approach, and the risk-neutral expectation approach. In both cases, the real-world drift, μ , of the share price does not matter. Not surprisingly, a theorem from pure mathematics, the Feynman-Kac theorem, joins the two approaches together by stating that certain second-order linear partial differential equations can be solved by taking expectations of diffusive processes.

2.4 Beyond Black-Scholes

For a number of reasons, the theory outlined above is not the end of the story. There is considerable evidence that the log of the share price does not follow a Brownian motion with drift. In particular, market crashes occur. For example, in October 1987 the stock market fell by 30% in one day and financial institutions found that their replication strategies failed badly. Mathematically, a crash corresponds to a jump in the share price, and Brownian motion has the property that all paths are continuous. Thus the Black-Scholes model failed to capture an important feature of share-price evolution.

A reflection of this failure is that options on the same share but with differing strike prices often trade with different volatilities, despite the fact that the BS model suggests that all options should trade with the same volatility. The graph of volatility as a function of the strike price is normally in the shape of a smile, displaying the disbelief of traders in the Black-Scholes model.

Another deficiency of the model is that it assumes that the volatility is constant. In practice, market activity varies in intensity and goes through some periods

Terri: Tim would like to keep the heading as it is. OK?

when share prices are much more volatile and others when they are much less so. Models must therefore be corrected to take account of the stochasticity of volatility, and the prediction of volatility over the life of an option is an important part of its pricing. Such models are called stochastic volatility models.

If one examines the data on small-scale share movements, one quickly discovers that they do not resemble a diffusion. They appear to be more like a series of small jumps than a Brownian motion. However, if one rescales time so that it is based on the number of trades that have occurred rather than on calendar time, then the returns do become approximately normal. One way to generalize the Black-Scholes model is to introduce a second process that expresses trading time. An example of such a model is known as the *variance gamma model*. More generally, the theory of Lévy processes has been applied to develop wider theories of price movements for shares and other assets.

Most generalizations of the Black-Scholes model do not retain the property of market completeness. They therefore give rise to many prices for options rather than just one.

2.5 Exotic Options

Many derivatives have quite complicated rules to determine their payoffs. For example, a *barrier option* can be exercised only if the share price does not go below a certain level at any time during the contract's life, and an *Asian option* pays a sum that depends on the average of the share price over certain dates rather than on the price at expiry. Or the derivative might depend upon several assets at once, such as, for example, the right to buy or sell a basket of shares for a certain price. It is easy to write down expressions for the value of such derivatives in the Black-Scholes model, either via a PDE or as a risk-neutral expectation. It is not so easy to evaluate these expressions. Much research is therefore devoted to developing efficient methods of pricing such options. In certain cases it is possible to develop analytic expressions. However, these tend to be the exception rather than the rule, and this means that one must resort to numerical techniques.

There is a wealth of methods for solving PDEs and these can be applied to derivatives-pricing problems. One difficulty in mathematical finance, however, is that the PDE can be very high dimensional. For example, if one is trying to evaluate a credit product depending on 100 assets, the PDE could be 100 dimensional. PDE

methods are most effective for low-dimensional problems, and so research is devoted to trying to make them effective in a wider range of cases.

One method that is less affected by dimensionality is Monte Carlo evaluation. The basis of this method is very simple: both intuitively and (via the law of large numbers) mathematically, an expectation is the long-run average of a series of independent samples of a random variable X . This immediately yields a numerical method for estimating $\mathbb{E}(f(X))$. One simply takes many independent samples X_i of X , calculates $f(X_i)$ for each one and computes their average. It follows from the CENTRAL LIMIT THEOREM [III.73 §5] that the error after N draws is approximately distributed as a normal distribution with variance equal to $N^{-1/2}$ times the variance of $f(X)$. The rate of convergence is therefore dimension independent. If the variance of $f(X)$ is large, it may still be rather slow, however. Much effort is therefore devoted by financial mathematicians to developing methods of reducing the variance when one computes high-dimensional integrals.

2.6 Vanilla versus Exotics

Generally, a simple option to buy or sell an asset is known as a *vanilla option*, whereas a more complicated derivative is known as an *exotic option*. An essential difference between the pricing of the two is that one can hedge an exotic option not just with the underlying share, but also by trading appropriately in the vanilla options on that share. Typically, the price of a derivative will depend not just on observable inputs, such as the share price and interest rates, but also on unobservable parameters, such as the volatility of the share price or the frequency of market crashes, which cannot be measured but only estimated.

When trading exotic options, one wishes to reduce dependence upon these unobservable inputs. A standard way to do this is to trade vanilla options in such a way as to make the rate of change of the value of the portfolio with respect to such parameters equal to zero. A small misestimation of their value will then have little effect on the worth of the portfolio.

This means that when one prices exotic options, one wishes not just to capture the dynamics of the underlying asset accurately but also to price all the vanilla options on that asset correctly. In addition, the model will predict how the prices of vanilla options change when the share price changes. We want these predictions to be accurate.

The BS model takes volatility to be constant. However, one can modify it so that the volatility varies with the share price and over time. One can choose how it varies in such a way that the model matches the market prices of all vanilla options. Such models are known as *local volatility models* or *Dupire models*. Local volatility models were very popular for a while, but have become less so because they give a poor model for how the prices of vanilla options change over time.

Much of the impetus behind the development of the models we mentioned in section 2.4 comes from the desire to produce a model that is computationally tractable, prices all vanilla options correctly, and produces realistic dynamics for both the underlying assets and the vanilla options. This problem has still not been wholly solved. There tends to be a trade-off between realistic dynamics and perfect matching of the vanilla options market. One compromise is to fit the market as well as possible using a realistic model and then to superimpose a local volatility model to remove the remaining errors.

3 Risk Management

3.1 Introduction

Once we have accepted that it is impossible to make money in finance without taking risk, it becomes important to be able to measure and quantify risks. We wish to measure accurately how much risk we are taking and decide whether we are comfortable with that level of risk. For a given level of risk, we want to maximize our expected return. When considering a new transaction, we will want to examine how it affects our risk levels and returns. Certain transactions may even reduce our risk while increasing our returns if they cancel out other risk. (A risk that can be canceled out by other risks that have a tendency to move in the opposite direction is called *diversifiable*.)

The control of risk becomes particularly important when dealing with portfolios of derivatives, which are often of zero value initially but which can very quickly change value. Placing a limit on the value of the contracts held is therefore not of much use, and controls based on deal sizes are complicated by the fact that often many derivatives contracts largely cancel each other out; it is the *residual* risk that one wishes to control.

3.2 Value-at-Risk

One method of limiting an institution's risks in derivatives trading is to place a limit on the amount it can lose with a given probability over a specified period of time. For instance, one might consider the losses at a 1% level over ten days, or at a 5% level over one day. This value is called *Value-at-Risk* or VAR.

To compute VAR one has to build up a probabilistic model of how the portfolio of derivatives might change in value over the time period. This requires a model of how all the underlying assets can move. Given this model, one then builds up the distribution of possible profits and losses over the given time period. Once one has this distribution one simply reads off the desired percentile.

The issues involved in modeling the changes for VAR computation are quite different from those for derivatives pricing. Typically, a VAR computation is done over a very short time period, such as one or ten days, unlike the pricing of an option, which deals with a long time frame. Also, one is not interested in the typical path for VAR, but instead one focuses on the extreme moves. In addition, since it is the VAR of an entire portfolio that matters, one has to develop an accurate model of the underlying assets' *joint* distributions: the movement of one underlying asset could magnify the price movement of another, or it could act as a hedge.

There are two main approaches to developing a probabilistic model for computing VAR. The first, the historical approach, is to record all the daily changes over some time period, for example two years, and then assume that the set of changes tomorrow will be identical to one of the sets of changes we have recorded. If we assign equal probability to each of those changes, then we get an approximation to the profit and loss distribution, from which we can read off the desired percentile. Note that as we are using a day's change for all assets simultaneously, we automatically get an approximation to the joint distribution of all the asset prices.

A second approach is to assume that asset price movements come from some well-known class of distributions. For example, we could assume that the logs of the asset price movements are jointly normal. We would then use historical data to estimate the volatilities and the correlations between the various prices. The main difficulty with this approach is obtaining robust estimates of the correlations given a limited amount of data.

4 Portfolio Optimization

4.1 Introduction

The job of a fund manager is to maximize the return on the money invested while minimizing the risk. If we assume that markets are efficient, then there is no point in trying to pick shares that we believe to be undervalued as we have assumed that they do not exist. A corollary is that just as no shares are good buys, no shares are bad buys. In any case, over half the shares in the market are owned via funds and therefore under the control of fund managers. Therefore, the average fund manager cannot expect to outperform the market.

It may seem that this does not leave much for fund managers to do, but in fact it leaves two things.

- (i) They can attempt to control the amount of risk they are taking.
- (ii) For a given level of risk, they can maximize their expected return.

To do these things requires an accurate model of the joint distribution of asset prices over the longer term, and a quantifiable notion of risk.

4.2 The Capital Asset Pricing Model

Portfolio theory has been in its modern form for longer than derivatives pricing. As an area, it relies less on stochastic calculus and more on economics. We briefly review the key ideas. The best-known model for modeling portfolio returns is the *capital asset pricing model* (or CAPM), which was introduced in the 1950s by Sharpe (see Sharpe 1964), and is still ubiquitous. Sharpe's model built on earlier work of Markowitz (1952).

The fundamental problem in this area is to assess what portfolio of assets, generally shares, an investor should hold in order to maximize returns at a given level of risk. The theory requires assumptions to be made about the joint distribution of share returns, e.g., joint normality, and/or about the risk preferences of investors, e.g., that they only care about the mean and variance of returns.

Under these assumptions, the CAPM yields the result that every investor should hold a multiple of the "market portfolio," which is essentially a portfolio consisting of everything traded in appropriate quantities to achieve maximum diversification, together with a certain amount of the risk-free asset. The relative amounts are determined by the investor's risk preferences.

A consequence of the model is the distinction between diversifiable risk and undiversifiable risk. While investors are compensated for taking undiversifiable, or systematic, risk via higher expected returns, diversifiable risk does not carry a risk premium. This is because one can cancel out diversifiable risk by holding appropriate combinations of other assets. Therefore, if it carried a risk premium, investors could receive extra return without taking any risk.

Much of the current research in this area is directed at trying to find more accurate models for the joint distribution of returns, and at finding techniques that estimate the parameters of such returns. A related problem is the "equity premium puzzle," which is that the excess return on investing in shares is much higher than the model predicts for reasonable levels of risk aversion.

5 Statistical Arbitrage

We only briefly mention statistical arbitrage as it is a rapidly changing area that is shrouded in secrecy. The fundamental idea in this area is to squeeze information out of asset price movements that the market has not already acted on. It therefore contradicts the principle of market efficiency, which says that all available information is already encoded in the market price. One explanation is that it is the action of taking such arbitrages that makes the market efficient.

Further Reading

- Bachelier, L. 1900. *La Théorie de la Spéculation*. Paris: Gauthier-Villars.
- Black, F., and M. Scholes. 1973. The valuation of options and corporate liabilities. *Journal of Political Economy* 81: 637–54.
- Einstein, A. 1985. *Investigations on the Theory of the Brownian Movement*. New York: Dover.
- Harrison, J. M., and D. M. Kreps. 1979. Martingales and arbitrage in multi-period securities markets. *Journal of Economic Theory* 20:381–408.
- Harrison, J. M., and S. R. Pliska. 1981. Martingales and stochastic integration in the theory of continuous trading. *Stochastic Processes and Applications* 11:215–60.
- Markowitz, H. 1952. Portfolio selection. *Journal of Finance* 7:77–99.
- Sharpe, W. 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance* 19:425–42.

VII.10 Mathematical Statistics

Persi Diaconis

1 Introduction

Suppose you want to measure something: your height, or the velocity of an airplane for example. You take repeated measurements x_1, x_2, \dots, x_n and you would like to combine them into a final estimate. An obvious way of doing this is to use the *sample mean* $(x_1 + x_2 + \dots + x_n)/n$. However, modern statisticians use many other estimators, such as the median or the *trimmed mean* (where you throw away the largest and smallest 10% of the measurements and take the average of what is left). Mathematical statistics helps us to decide when one estimate is preferable to another. For example, it is intuitively clear that throwing away a random half of the data and averaging the rest is foolish, but setting up a framework that shows this clearly turns out to be a serious enterprise. One benefit of the undertaking is the discovery that the mean turns out to be inferior to nonintuitive “shrinkage estimators” even when the data are drawn from a PROBABILITY DISTRIBUTION [III.73] as natural as the bell-shaped curve (that is, are NORMALLY DISTRIBUTED [III.73 §5]).

To get an idea of why the mean may not always give you the most useful estimate, consider the following situation. You have a collection of a hundred coins and you would like to estimate their biases. That is, you would like to estimate a sequence of a hundred numbers, where the n th number θ_n is the probability that the n th coin will come up heads when it is flipped. Suppose that you flip each coin five times and note down how many times it shows heads. What should your estimate be for the sequence $(\theta_1, \dots, \theta_{100})$? If you use the means, then your guess for θ_n will be the number of times the n th coin shows heads, divided by 5. However, if you do this, then you are likely to get some very anomalous results. For instance, if all the coins happen to be unbiased, then the probability that any given coin shows up heads five times is $1/32$, so you are likely to guess that around three of the coins have biases of 1. So you will be guessing that if you flip those coins five hundred times then they will come up heads every single time.

Many alternative methods of estimation have been proposed in order to deal with this obvious problem. However, one must be careful: if a coin comes up heads five times it could be that θ_i really is equal to 1. What reason is there to believe that a different method of estimation is not in fact taking us further from the truth?

Here is a second example, drawn from work of Bradley Efron, this time concerning a situation from real life. Table 1 shows the batting averages of eighteen baseball players. The first column shows the proportion of “hits” for each player in their first forty-five times at bat, and the second column shows the proportion of hits at the end of the season. Consider the task of predicting the second column given only the first column. Once again, the obvious approach is to use the average. In other words, one would simply use the first column as a predictor of the second column. The third column is obtained by a shrinkage estimator: more precisely, it takes a number y in the first column and replaces it by $0.265 + 0.212(y - 0.265)$. The number 0.265 is the average of the entries in the first column, so the shrinkage estimator is replacing each entry in the first column by one that is about five times closer to the average. (How the number 0.212 is chosen will be explained later.) If you look at the table, you will see that the shrinkage estimators in the third column are better predictors of the second column in almost every case, and certainly on average. Indeed, the sum of squared differences between the James–Stein estimator and the truth divided by the sum of squared differences between the usual estimator and the truth is 0.29. That is a threefold improvement.

There is beautiful mathematics behind this improvement and a clear sense in which the new estimator is *always* better than the average. We describe the framework, ideas, and extensions of this example as an introduction to the mathematics of statistics.

Before beginning, it will be useful to distinguish between probability and statistics. In probability theory, one begins with a set X (for the moment taken to be finite) and a collection of numbers $P(x)$, one for each $x \in X$, which are positive and sum to one. This function $P(x)$ is called a *probability distribution*. The basic problem of probability is this. You are given the probability distribution $P(x)$ and a subset $A \subset X$, and you must compute or approximate $P(A)$, which is defined to be the sum of $P(x)$ for x in A . (In probabilistic terms, each x has a probability $P(x)$ of being chosen, and $P(A)$ is the probability that x belongs to A .) This simple formulation hides wonderful mathematical problems. For example, X might be the set of all sequences of pluses and minuses of length 100 (e.g., $+-+-+-----$), and each pattern might be equally likely, in which case $P(x) = 1/2^{100}$ for every sequence x . Finally, A might be the set of sequences such that for every positive integer $k \leq 100$ the number of $+$ symbols in the first k places

Table 1 Batting averages for
18 major league players in 1970.

Player number	Batting average after 45 at bats	Batting average remainder of season	James-Stein estimator	Remaining at bats
1	0.400	0.346	0.293	367
2	0.378	0.298	0.289	426
3	0.356	0.276	0.284	521
4	0.333	0.221	0.279	276
5	0.311	0.273	0.275	418
6	0.311	0.270	0.275	467
7	0.289	0.263	0.270	586
8	0.267	0.210	0.265	138
9	0.244	0.269	0.261	510
10	0.244	0.230	0.261	200
11	0.222	0.264	0.256	277
12	0.222	0.256	0.256	270
13	0.222	0.304	0.256	434
14	0.222	0.264	0.256	538
15	0.222	0.226	0.256	186
16	0.200	0.285	0.251	558
17	0.178	0.319	0.247	405
18	0.156	0.200	0.242	70

is larger than the number of $-$ symbols in the first k places. This is a mathematical model for the following probability problem: if you and a friend flip a fair coin a hundred times, then what is the chance that your friend is always ahead? One might expect this chance to be very small. It turns out, however, to be about $\frac{1}{12}$, though verifying this is a far from trivial exercise. (Our poor intuitions about chance fluctuations have been used to explain road rage: suppose you choose one of two lines at a toll booth. As you wait, you notice whether your line or the other has made more progress. We feel it should all balance out, but the calculations above show that a fair proportion of the time you are always behind—and *frustrated!*)

2 The Basic Problem of Statistics

Statistics is a kind of opposite of probability. In statistics, we are given a *collection* of probability distributions $P_\theta(x)$, indexed by some parameter θ . We see just one x and are required to guess which member of the family (which θ) was used to generate x . For example, let us keep X as the sequence of pluses and minuses of length 100, but this time let $P_\theta(x)$ be the chance of obtaining the sequence x if the probability of a plus is

θ and the probability of a minus is $1 - \theta$, with all terms in the sequence chosen independently. Here $0 \leq \theta \leq 1$, and $P_\theta(x)$ is easily seen to be $\theta^S(1 - \theta)^T$, where S is the number of times “+” appears in the sequence x and $T = 100 - S$ is the number of times “−” appears. This is a mathematical model for the following enterprise. You have a biased coin with a probability θ of turning up heads, but you do not know θ . You flip the coin a hundred times, and are required to estimate θ based on the outcome of the flips.

In general, for each $x \in X$, we want to find a guess, which we denote by $\hat{\theta}(x)$, for the parameter θ . That is, we want to come up with a function $\hat{\theta}$, which will be defined on the observation space X . Such functions are called *estimators*. The above simple formulation hides a wealth of complexity, since both the observation space X and the space Θ of possible parameters may be infinite, or even infinite dimensional. For example, in nonparametric statistics, Θ is often taken as the set of all probability distributions on X . All of the usual problems of statistics—design of experiments, testing hypotheses, prediction, and many others—fit into this framework. We will stick with the imagery of estimation.

To evaluate and compare estimators, one more ingredient is needed: you have to know what it means to get the right answer. This is formalized through the notion of a *loss function* $L(\theta, \hat{\theta}(x))$. One can think of this in practical terms: wrong guesses have financial consequences, and the loss function is a measure of how much it will cost if θ is the true value of the parameter but the statistician's guess is $\hat{\theta}(x)$. The most widely used choice is the *squared error* $(\theta - \hat{\theta}(x))^2$, but $|\theta - \hat{\theta}(x)|$ or $|\theta - \hat{\theta}(x)|/\theta$ and many other variants are also used. The *risk function* $R(\theta, \hat{\theta})$ measures the expected loss if θ is the true parameter and the estimator $\hat{\theta}$ is used. That is,

$$R(\theta, \hat{\theta}) = \int L(\theta, \hat{\theta}(x))P_\theta(dx).$$

Here, the right-hand side is notation for the average value of $L(\theta, \hat{\theta}(x))$ if x is chosen randomly according to the probability distribution P_θ . In general, one would like to choose estimators that will make the risk function as small as possible.

3 Admissibility and Stein's Paradox

We now have the basic ingredients: a family $P_\theta(x)$ and a loss function L . An estimator $\hat{\theta}$ is called *inadmissible*

if there is a better estimator θ^* , in the sense that

$$R(\theta, \theta^*) < R(\theta, \hat{\theta}) \quad \text{for all } \theta.$$

In other words, the expected loss with θ^* is less than the expected loss with $\hat{\theta}$, whatever the true value of θ .

Given our assumptions (the model P_θ and loss function L) it seems silly to use an inadmissible estimator. However, one of the great achievements of mathematical statistics is Charles Stein's proof that the usual least-squares estimator, which does not at first glance seem silly at all, is inadmissible in natural problems. Here is that story.

Consider the basic measurement model

$$X_i = \theta + \epsilon_i, \quad 1 \leq i \leq n.$$

Here X_i is the i th measurement, θ is the quantity to be estimated, and ϵ_i is measurement error. The classical assumptions are that the measurement errors are independently and normally distributed: that is, they are distributed according to the bell-shaped, or Gaussian, curve $e^{-x^2/2}/\sqrt{2\pi}$, $-\infty < x < \infty$. In terms of the language we introduced earlier, the measurement space X is \mathbb{R}^n , the parameter space θ is \mathbb{R} , and the observation $x = (x_1, x_2, \dots, x_n)$ has probability density $P_\theta(x) = \exp[-\frac{1}{2} \sum_1^n (x_i - \theta)^2] / (\sqrt{2\pi})^n$. The usual estimator is the mean: that is, if $x = (x_1, \dots, x_n)$, then one takes $\hat{\theta}(x)$ to be $(x_1 + \dots + x_n)/n$. It has been known for a long time that if the loss function $L(\theta, \hat{\theta}(x))$ is defined to be $(\theta - \hat{\theta}(x))^2$, then the mean is an admissible estimator. It has many other optimal properties as well (for example, it is the best linear unbiased estimator, and it is minimax—a property that will be defined later in this article).

Now suppose that we wish to estimate *two* parameters, θ_1 and θ_2 , say. This time we have two sets of observations, X_1, \dots, X_n and Y_1, \dots, Y_m , with $X_i = \theta_1 + \epsilon_i$ and $Y_j = \theta_2 + \eta_j$. The errors ϵ_i and η_j are independent and normally distributed, as above. The loss function $L((\theta_1, \theta_2), (\hat{\theta}_1(x), \hat{\theta}_2(y)))$ is now defined to be $(\theta_1 - \hat{\theta}_1(x))^2 + (\theta_2 - \hat{\theta}_2(y))^2$: that is, you add up the squared errors from the two parts. Again, the mean of the X_i and the mean of the Y_i make up an admissible estimator for (θ_1, θ_2) .

Consider the same setup with three parameters, $\theta_1, \theta_2, \theta_3$. Again, $X_i = \theta_1 + \epsilon_i$, $Y_i = \theta_2 + \eta_j$, $Z_k = \theta_3 + \delta_k$ are independent and all the error terms are normally distributed. Stein's surprising result is that for three

(or more) parameters the estimator

$$\begin{aligned} \hat{\theta}_1(x) &= (x_1 + \dots + x_n)/n, \\ \hat{\theta}_2(y) &= (y_1 + \dots + y_m)/m, \\ \hat{\theta}_3(z) &= (z_1 + \dots + z_l)/l \end{aligned}$$

is *inadmissible*: there are other estimators that do better in all cases. For example, if p is the number of parameters (and $p \geq 3$), then the *James-Stein estimator* is defined to be

$$\hat{\theta}_{js} = \left(1 - \frac{p-2}{\|\hat{\theta}\|}\right)_+ \hat{\theta}.$$

Here we are using the notation X_+ to denote the maximum of X and 0; θ stands for the vector $(\theta_1, \dots, \theta_p)$ of all the averages and $\|\hat{\theta}\|$ is notation for $(\theta_1^2 + \dots + \theta_p^2)^{1/2}$.

The James-Stein estimator satisfies the inequality $R(\theta, \hat{\theta}_{js}) < R(\theta, \hat{\theta})$ for all θ , and therefore the usual estimator $\hat{\theta}$ is indeed inadmissible. The James-Stein estimator shrinks the classical estimator toward zero. The amount of shrinkage is small if $\|\hat{\theta}\|^2$ is large and appreciable for $\|\hat{\theta}\|^2$ near zero. Now the problem as we have described it is invariant under translation, so if we can improve the classical estimate by shrinking toward zero, then we must be able to improve it by shrinking toward any other point. This seems very strange at first, but one can obtain some insight into the phenomenon by considering the following informal description of the estimator. It makes an a priori guess θ_0 at θ . (This guess was zero above.) If the usual estimator $\hat{\theta}$ is close to the guess, in the sense that $\|\hat{\theta}\|$ is small, then it moves $\hat{\theta}$ toward the guess. If $\hat{\theta}$ is far from the guess, it leaves $\hat{\theta}$ alone. Thus, although the estimator moves the classical estimator toward an arbitrary guess, it does so only if there are reasons to believe that the guess is a good one. With four or more parameters the data can in fact be used to suggest which point θ_0 one should use as the initial guess. In the example of table 1, there are eighteen parameters, and the initial guess θ_0 was the constant vector with all its eighteen coordinates equal to the average 0.265. The number 0.212 that was used for the shrinking is equal to $1 - 16/\|\theta - \theta_0\|$. (Note that for this choice of θ_0 , $\|\theta - \theta_0\|$ is the standard deviation of the parameters that make up θ .)

The mathematics used to prove inadmissibility is an elegant blend of harmonic function theory and tricky calculus. The proof itself has had many ramifications: it gave rise to what is called "Stein's method" in probability theory—this is a method for proving things like the central limit theorem for complex dependent problems. The mathematics is "robust," since it is applicable

to nonnormal error distributions, a variety of different loss functions, and estimation problems far from the measurement model.

The result has had enormous practical application. It is routinely used in problems where many parameters have to be simultaneously estimated. Examples include national laboratories' estimates of the percentage of defectives when they are looking at many different products at once, and the simultaneous estimate of census undercounts for each of the fifty states in the United States. The apparent robustness of the method is very useful for such applications: even though the James-Stein estimator was derived for the bell-shaped curve, it seems to work well, without special assumptions, in problems where its assumptions hold only roughly. Consider the baseball players above, for example. Adaptations and variations abound. Two popular ones are called empirical Bayes estimates (now widely used in genomics) and hierarchical modeling (now widely used in the assessment of education).

The mathematical problems are far from completely solved. For example, the James-Stein estimator is itself inadmissible. (It can be shown that any admissible estimator in a normal measurement problem is an analytic function of the observations. The James-Stein estimator is, however, clearly not analytic because it involves the nondifferentiable function $x \mapsto x_+$.) While it is known that there is little practical improvement possible, the search for an admissible estimator that is always better than the James-Stein estimator is a tantalizing research problem.

Another active area of research in modern mathematical statistics is to understand which statistical problems give rise to Stein's paradox. For example, although at the beginning of this essay we discussed some inadequacies of the usual maximum-likelihood estimator for estimating the biases of a hundred coins, it turns out that that estimator is admissible! In fact, the maximum-likelihood estimator is admissible for any problem with finite state spaces.

4 Bayesian Statistics

The *Bayesian approach* to statistics adds one further ingredient to the family P_θ and loss function L . This is known as a *prior probability distribution* $\pi(\theta)$, which gives different weights to different values of the parameter θ . There are many ways of generating a prior distribution: it may quantify the working scientists' best guess at θ ; it may be derived from previous studies

or estimates; or it may just be a convenient way to generate estimators. Once the prior distribution $\pi(\theta)$ has been specified, the observation x and Bayes's theorem combine to give a *posterior distribution* for θ , here denoted $\pi(\theta|x)$. Intuitively, if x is your observation, then $\pi(\theta|x)$ measures how likely it is that θ was the parameter, given that the parameter was generated from the probability distribution π . The mean value of θ with respect to the posterior distribution $\pi(\theta|x)$ gives a *Bayes estimator*:

$$\hat{\theta}_{\text{Bayes}}(x) = \int \theta \pi(\theta|x).$$

For the squared-error loss function, all Bayes estimators are admissible, and, in the converse direction, any admissible estimator is a limit of Bayes estimators. (However, not every limit of Bayes estimators is admissible: indeed, the average, which we have seen to be inadmissible, is a limit of Bayes rules.) The point for the present discussion is this. In a wide variety of practical variations of the measurement problem—things like regression analysis or the estimation of correlation matrices—it is relatively straightforward to write down sensible Bayes estimators that incorporate available prior knowledge. These estimators include close cousins of the James-Stein estimator, but they are more general, and allow it to be routinely extended to almost any statistical problem.

Because of the high-dimensional integrals involved, Bayes estimates can be difficult to compute. One of the great advances in this area is the use of computer-simulation algorithms, called variously *Markov chain Monte Carlo* or *Gibbs samplers*, to compute useful approximations to Bayes estimators. The whole package—provable superiority, easy adaptability, and ease of computation—has made this Bayesian version of statistics a practical success.

5 A Bit More Theory

Mathematical statistics makes good use of a wide range of mathematics: fairly esoteric analysis, logic, combinatorics, algebraic topology, and differential geometry all play a role. Here is an application of group theory. Let us return to the basic setup of a sample space X , a family of probability distributions $P_\theta(x)$, and a loss function $L(\theta, \hat{\theta}(x))$. It is natural to consider how the estimator changes when you change the units of the problem: from pounds to grams, or from centimeters to inches, say. Will this have a significant impact on the mathematics? One would expect not, but if we want to

think about this question precisely then it is useful to consider a group G of transformations of X . For example, linear changes of units correspond to the *affine group*, which consists of transformations of the form $x \mapsto ax + b$. The family $P_\theta(x)$ is said to be *invariant* under G if for each element g of G the transformed distribution $P_\theta(xg)$ is equal to a distribution $P_{\hat{\theta}}(x)$ for some other $\hat{\theta}$ in Θ . For example, the family of normal distributions

$$\frac{\exp\left[-\frac{\frac{1}{2}(x - \theta_1)^2}{2\theta_2^2}\right]}{\sqrt{2n\theta_2^2}}, \quad -\infty < \theta_1 < \infty, \quad 0 < \theta_2 < \infty,$$

is invariant under $ax+b$ transformations: if you change x to $ax + b$, then after some easy manipulations you can rewrite the resulting modified formula in the form $\exp[-\frac{1}{2}(x - \phi_1)^2/2\phi_2^2]/\sqrt{2n\phi_2^2}$ for some new parameters ϕ_1 and ϕ_2 . An estimator $\hat{\theta}$ is called *equivariant* if $\hat{\theta}(xg) = \hat{\theta}(x)$. This is a formal way of saying that if you change the data from one unit to another, then the estimate transforms as it should. For example, suppose your data are temperatures presented in centigrade and you want an answer in Fahrenheit. If your estimator is equivariant, then it will make no difference whether you first apply the estimator and then convert the answer into Fahrenheit or first convert all the data into Fahrenheit and then apply the estimator.

The multivariate normal problem that underlies Stein's paradox is invariant under a variety of groups, including the p -dimensional group of Euclidean motions (rotations and translations). However, the James-Stein estimator is not equivariant, since, as we have already discussed, it depends on the choice of origin. This is not necessarily bad, but it is certainly thought provoking. If you ask a working scientist if they want a "most accurate" estimator, they will say "of course." If you ask if they insist on equivariance, "of course" will follow as well. One way of expressing Stein's paradox is the statement that the two desiderata—accuracy and invariance—are *incompatible*. This is one of many places where mathematics and statistics part company. Deciding whether mathematically optimal procedures are "sensible" is important and hard to mathematize.

Here is a second use of group theory. An estimator $\hat{\theta}$ is called *minimax* if it minimizes the maximum risk over all θ . Minimax corresponds to playing things safe: you have optimal behavior (that is, the least possible risk) in the worst case. Finding minimax estimators in natural problems is hard, honest work. For example,

the vector of means is a minimax estimator in normal location problems. The work is easier if the problem is invariant under a group. Then one can first search for best invariant estimators. Invariance often reduces things to a straightforward calculus problem. Now the question arises of whether an estimator that is minimax among invariant estimators is minimax among all estimators. A celebrated theorem of Hurt and Stein says "yes" if the group involved is nice (e.g., Abelian or compact or amenable). Determining whether the best invariant estimator is minimax when the group is not nice is a challenging open problem in mathematical statistics. And it is not just a mathematical curiosity. For example, the following problem is very natural, and invariant under the group of invertible matrices: given a sample from the multivariate normal distribution, estimate its correlation matrix. In this case, the group is not nice and good estimates are not known.

6 Conclusion

The point of this article is to show how mathematics enters and enriches statistics. To be sure, there are parts of statistics that are hard to mathematize: graphical displays of data are an example. Further, much of modern statistical practice is driven by the computer. There is no longer any need to restrict attention to tractable families of probability distributions. Complex and more realistic models can be used. This gives rise to the subject of statistical computing. Nonetheless, every once in a while someone has to think about what the computer *should* do and determine whether one innovative procedure works better than another. Then, mathematics holds its own. Indeed, mathematizing modern statistical practice is a challenging, rewarding enterprise, of which Stein's estimator is a current highlight. This endeavor gives us something to aim for and helps us to calibrate our day-to-day achievements.

Further Reading

- Berger, J. O. 1985. *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. New York: Springer.
- Lehmann, E. L., and G. Casella. 2003. *Theory of Point Estimation*. New York: Springer.
- Lehmann, E. L., and J. P. Romano. 2005. *Testing Statistical Hypotheses*. New York: Springer.
- Schervish, M. 1996. *Theory of Statistics*. New York: Springer.

VII.11 Mathematics and Medical Statistics

David J. Spiegelhalter

1 Introduction

There are many ways in which mathematics has been applied in medicine: for example, the use of differential equations in pharmacokinetics and models for epidemics in populations; and FOURIER ANALYSIS [III.27] of biological signals. Here we are concerned with medical statistics, by which we mean collecting data about individuals and using it to draw conclusions about the development and treatment of disease. This definition may appear to be rather restrictive, but it includes all of the following: randomized clinical trials of therapies, evaluating interventions such as screening programs, comparing health outcomes in different populations and institutions, describing and comparing the survival of groups of individuals, and modeling the way in which a disease develops, both naturally and when it is influenced by an intervention. In this article we are not concerned with *epidemiology*, the study of why diseases occur and how they spread, although most of the formal ideas described here can be applied to it.

After a brief historical introduction, we shall summarize the varied approaches to probabilistic modeling in medical statistics. We shall then illustrate each one in turn using data about the survival of a sample of patients with lymphoma, showing how alternative “philosophical” perspectives lead directly to different methods of analysis. Throughout, we shall give an indication of the mathematical background to what can appear to be a conceptually untidy subject.

2 A Historical Perspective

One of the first uses of probability theory in the late seventeenth century was in the development of “life-tables” of mortality in order to decide premiums for annuities, and Charles Babbage’s work on life-tables in 1824 helped motivate him to design his “difference engine” (although it was not until 1859 that Scheutz’s implementation of the engine finally calculated a life-table). However, statistical analysis of medical data was a matter of arithmetic rather than mathematics until the growth of the “biometric” school founded by Francis Galton and Karl Pearson at the end of the nineteenth century. This group introduced the use of families of

PROBABILITY DISTRIBUTIONS [III.73] to describe populations, as well as concepts of correlation and regression in anthropology, biology, and eugenics. Meanwhile, agriculture and genetics motivated Fisher’s huge contributions in the theory of likelihood (see below) and significance testing. Postwar statistical developments were influenced by industrial applications and a U.S.-led increase in mathematical rigor, but from around the 1970s medical research, particularly concerning randomized trials and survival analysis, has been a major methodological driver in statistics.

For around thirty years after 1945 there were repeated attempts to put statistical inference on a sound foundational or axiomatic basis, but no consensus could be reached. This has given rise to a widespread ecumenical perspective which makes use of a mix of statistical “philosophies” which we shall illustrate below. The somewhat uncomfortable lack of an axiomatic basis can make statistical work deeply unattractive to many mathematicians, but it provides a great stimulus to those engaged in the area.

3 Models

In this context, by a *model* we mean a mathematical description of a probability distribution for one or more currently uncertain quantities. Such a quantity might, for example, be the outcome of a patient who is treated with a particular drug, or the future survival time of a patient with cancer. We can identify four broad approaches to modeling—these brief descriptions make use of terms that will be covered properly in later sections.

- (i) A *nonparametric* or “model-free” approach that leaves unspecified the precise form for the probability distributions of interest.
- (ii) A *full parametric model* in which a specific form is assumed for each probability distribution, which depends on a limited number of unknown parameters.
- (iii) A *semi-parametric* approach in which only part of the model is parametrized, while the rest is left unspecified.
- (iv) A *Bayesian* approach in which not only is a full parametric model specified, but an additional “prior” distribution is provided for the parameters.

These are not absolute distinctions: for example, some apparently “model-free” procedures may turn out to

match procedures that are derived under certain parametric assumptions.

Another complicating factor is the multiplicity of possible aims of a statistical analysis. These may include

- *estimating* unknown parameters, such as the mean reduction in blood pressure when giving a certain dose of a certain drug to a defined population;
- *predicting* future quantities, such as the number of people with AIDS in a country in ten years' time;
- *testing a hypothesis*, such as whether a particular drug improves survival for a particular class of patients, or equivalently assessing the “null hypothesis” that it has no effect;
- *making decisions*, such as whether to provide a particular treatment in a health care system.

A common aspect of these objectives is that any conclusion should be accompanied by some form of assessment of the potential for an error having been made, and any estimate or prediction should have an associated expression of uncertainty. It is this concern for “second-order” properties that distinguishes a statistical “inference” based on probability theory from a purely algorithmic approach to producing conclusions from data.

4 The Nonparametric or “Model-Free” Approach

Now let us introduce a running example that will be used to illustrate the various approaches.

Matthews and Farewell (1985) report data on sixty-four patients from the Fred Hutchinson Cancer Research Center in Seattle who had been diagnosed with advanced-stage non-Hodgkin's lymphoma: for each patient the information comprises their follow-up time since diagnosis, whether their follow-up ended in death, whether they presented with clinical symptoms, their stage of disease (stage IV or not), and whether a large abdominal mass (greater than 10 cm) was present. Such information has many uses. For example, we may wish to look at the general distribution of survival times, or assess which factors most influence survival, or provide a new patient with an estimate of their chance of surviving, say, five years. This is, of course, too small and limited a data set to draw firm conclusions, but it allows us to illustrate the different mathematical tools that can be used.

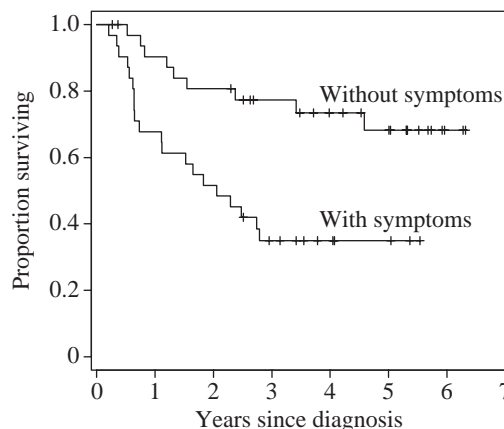


Figure 1 Kaplan-Meier nonparametric survival curves for lymphoma patients with and without clinical symptoms at diagnosis.

We need to introduce a few technical terms. Patients who are still alive at the end of data collection, or have been lost to follow-up, are said to have their survival times “censored”: all we know is that they survived beyond the last time that any data was recorded about them. We also tend to call times of death “failure” times, since the forms of analysis do not just apply to death. (This term also reflects the close connection between this area and *reliability theory*.)

The original approach to such survival data was “actuarial,” using the life-table techniques mentioned previously. Survival times are grouped into intervals such as years, and simple estimates are made of one's chance of dying in an interval given that one was alive at the start of it. Historically, this probability was known as the “force of mortality,” but now it is usually called the *hazard*. A simple approach like this may be fine for describing large populations.

It was not until Kaplan and Meier (1958) that this procedure was refined to take into account the exact rather than the grouped survival times: with over thirty thousand citations, their paper is one of the most referenced papers in all of science. Figure 1 shows so-called Kaplan-Meier curves for the groups of patients with ($n = 31$) and without ($n = 33$) clinical symptoms at diagnosis.

These curves represent estimates of the underlying *survival function*, whose value at a time t is thought of as the probability that a typical patient will survive until that time. The obvious way of producing such a curve is simply to let its value at time t be the proportion of

the initial sample that is still alive. However, this does not quite work, because of the censored patients. So instead, if a patient dies at time t , and if, just before time t , there are m patients still in the sample, then the value of the curve is multiplied by $(m - 1)/m$; and if a patient is censored then the value stays the same. (The tick marks on the curves show the censored survival times.) The set of patients alive just before time t is called the *risk set* and the hazard at t is estimated to be $1/m$. (We are assuming that two people do not die at the same time, but it is easy to drop that assumption and make appropriate adjustments.)

Although we do not assume that the actual survival curve has any particular functional form, we do need to make the qualitative assumption that the censoring mechanism is independent of the survival time. (For example, it is important that those who are about to die are not for some reason preferentially removed from the study.) We also need to provide error bounds on the curves: these can be based on a variance formula developed by Major Greenwood in 1926. ("Major" was his name rather than a title, one of the few characteristics he shared with Count Basie and Duke Ellington.)

The "true underlying survival curve" is a theoretical construct, and not something that one can directly observe. One can think of it as the survival experience that would be observed in a vast population of patients, or equivalently the expected survival for a new individual drawn at random from that population. As well as estimating these curves for the two groups of patients, we may wish to test hypotheses about them. A typical one would be that the true underlying survival curves in the two groups are precisely the same. Traditionally such "null" hypotheses are denoted H_0 , and the traditional way to test them is to determine how unlikely it is that we would observe two Kaplan-Meier curves that are so far apart if H_0 were true. One can construct a summary measure, known as a *test statistic*, that is large if the observed curves are very different. For example, one possibility is to contrast the observed number of deaths in those with symptoms ($O = 20$) with the number one would have expected if H_0 were true ($E = 11.9$). Under the null hypothesis it turns out there is only a 0.2% chance of observing such a high discrepancy between O and E , which casts considerable doubt on the null hypothesis in this case.

When constructing intervals around estimates and testing hypotheses we require approximate probability distributions for our estimates and test statistics. From a mathematical perspective the important theory

therefore concerns large-sample distributions of functions of random variables, largely developed in the early twentieth century. Theories for optimal hypothesis testing were developed by Neyman and Pearson in the 1930s: the idea is to maximize the "power" of a test to detect a difference, while at the same time making sure that the probability of wrongly rejecting a null hypothesis is less than some acceptable threshold such as 5% or 1%. This approach still finds a role in the design of randomized clinical trials.

5 Full Parametric Models

Clearly we do not actually believe that deaths can only occur at the previously observed survival times shown in the Kaplan-Meier curve, so it seems reasonable to investigate a fairly simple functional form for the true survival function. That is, we assume that the survival function belongs to some natural class of functions, each of which can be fully parametrized by a small number of parameters, collectively denoted by θ . It is θ that we are trying to discover (or rather estimate with a reasonable degree of confidence). If we can do so, then the model is fully specified and we can even extrapolate a certain amount beyond the observed data. We first relate the survival function and the hazard, and then illustrate how observed data can be used to estimate θ in a simple example.

We assume that an unknown survival time has a probability density $p(t|\theta)$; without getting into technical details, this essentially corresponds to assuming that $p(t|\theta) dt$ is the probability of dying in a small interval t to $t + dt$. Then the survival function, given a particular value of θ , is the probability of surviving beyond t : we denote it by $S(t|\theta)$. To calculate it, we integrate the probability density over all times greater than t . That is,

$$S(t|\theta) = \int_t^\infty p(x|\theta) dx = 1 - \int_0^t p(x|\theta) dx.$$

From this and THE FUNDAMENTAL THEOREM OF CALCULUS [I.3 §5.5] it follows that $p(t|\theta) = -dS(t|\theta)/dt$. The hazard function $h(t|\theta) dt$ is the risk of death in the small interval t to $t + dt$, conditional on having survived to time t . Using the laws of elementary probability we find that

$$h(t|\theta) = p(t|\theta)/S(t|\theta).$$

For example, suppose we assume an exponential survival function with mean survival time θ , so that the probability of surviving beyond time t is $S(t|\theta) = e^{-t/\theta}$.

The density is $p(t|\theta) = e^{-t/\theta}/\theta$. Therefore, the hazard function is a constant $h(t|\theta) = 1/\theta$, so that $1/\theta$ represents the mortality rate per unit of time. For instance, were the mean postdiagnosis survival to be $\theta = 1000$ days, an exponential model would imply a constant $1/1000$ risk of dying each day, regardless of how long the patient had already survived after diagnosis. More complex parametric survival functions allow hazard functions that increase, decrease, or have other shapes.

When it comes to estimating θ we need Fisher's concept of *likelihood*. This takes the probability distribution $p(t|\theta)$ but considers it as a function of θ rather than t , and hence for observed t allows us to examine plausible values of θ that "support" the data. The rough idea is that we multiply together the probabilities (or probability densities) of the observed events, assuming the value of θ . In survival analysis, observed and censored failure times make different contributions to this product: an observed time t contributes $p(t|\theta)$, while a censored time contributes $S(t|\theta)$. If, for example, we assume that the survival function is exponential, then an observed failure time contributes $p(t|\theta) = e^{-t/\theta}/\theta$, and a censored time contributes $S(t|\theta) = e^{-t/\theta}$. Thus, in this case the likelihood is

$$L(\theta) = \prod_{i \in \text{Obs}} \theta^{-1} e^{-t_i/\theta} \prod_{i \in \text{Cens}} e^{-t_i/\theta} = \theta^{-n_0} e^{-T/\theta}.$$

Here "Obs" and "Cens" indicate the sets of observed and censored failure times. We denote their sizes by n_0 and n_c , respectively, and we denote the total follow-up time $\sum_i t_i$ by T . For the group of thirty-one patients presenting with symptoms we have $n_0 = 20$ and $T = 68.3$ years: figure 2 shows both the likelihood and its logarithm

$$LL(\theta) = -T/\theta - n_0 \log \theta.$$

We note that the vertical axis for the likelihood is unlabeled since only relative likelihood is important. A *maximum-likelihood estimate* (MLE) $\hat{\theta}$ finds parameter values that maximize this likelihood or equivalently the log-likelihood. Taking derivatives of $LL(\theta)$ and equating to 0 reveals that $\hat{\theta} = T/n_{\text{Obs}} = 3.4$ years, which is the total follow-up time divided by the number of failures. Intervals around MLEs may be derived by directly examining the likelihood function, or by making a quadratic approximation around the maximum of the log-likelihood.

Figure 3 shows the fitted exponential survival curves: loosely, we have carried out a form of curve fitting by selecting the exponential curves that maximize the

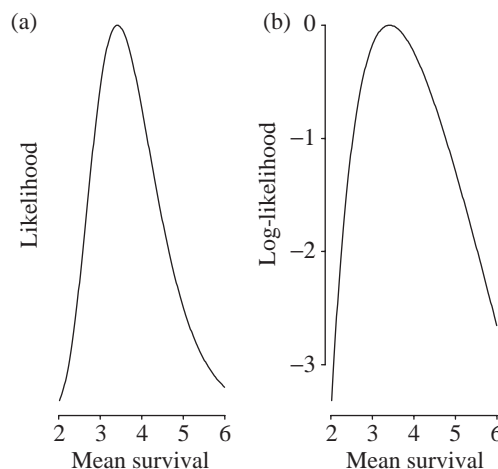


Figure 2 Likelihood and log-likelihood for mean survival time θ for lymphoma patients presenting with clinical symptoms.

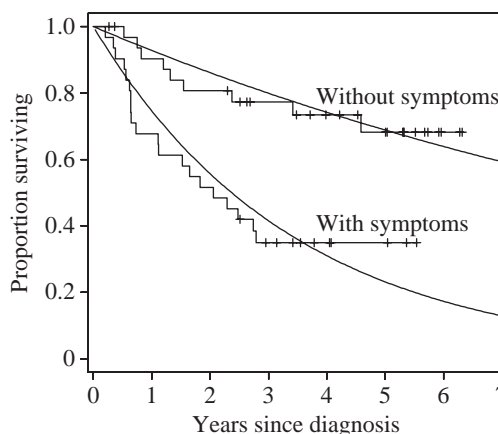


Figure 3 Fitted exponential survival curves for lymphoma patients.

probability of the observed data. Visual inspection suggests the fit may be improved by investigating a more flexible family of curves such as the Weibull distribution (a distribution widely used in reliability theory): to compare how well two models fit the data, one can compare their maximized likelihoods.

Fisher's concept of likelihood has been the foundation for most current work in medical statistics, and indeed statistics in general. From a mathematical perspective there has been extensive development relating the large-sample distributions of MLEs to the second derivative of the log-likelihood around its maxi-

mum, which forms the basis for most of the outputs of statistical packages. Unfortunately, it is not necessarily straightforward to scale up the theory to deal with multidimensional parameters. First, as likelihoods become more complex and contain increasing numbers of parameters, the technical problems of maximization increase. Second, the recurring difficulty with likelihood theory remains that of “nuisance parameters,” in which a part of the model is of no particular interest and yet needs to be accounted for. No generic theory has been developed, and instead there is a somewhat bewildering variety of adaptations of standard likelihood to specific circumstances, such as conditional likelihood, quasi-likelihood, pseudo-likelihood, extended likelihood, hierarchical likelihood, marginal likelihood, profile likelihood, and so on. Below we consider one extremely popular development, that of partial likelihood and the Cox model.

6 A Semi-Parametric Approach

Clinical trials in cancer therapy were a major motivating force in developing survival analysis—in particular, trials to assess the influence of a treatment on survival while taking account of other possible risk factors. In our simple lymphoma data set we have three risk factors, but in more realistic examples there will be many more. Fortunately, Cox (1972) showed that it was possible both to test hypotheses and to estimate the influence of possible risk factors, without having to go the whole way and specify the full survival function on the basis of possibly limited data.

The *Cox regression model* is based on assuming a hazard function of the form

$$h(t|\theta) = h_0(t)e^{\beta \cdot x}.$$

Here $h_0(t)$ is a *baseline hazard function* and β is typically a column vector of regression coefficients that measure the influence of a vector of risk factors x on the hazard. (The expression $\beta \cdot x$ denotes the scalar product of β and x .) The baseline hazard function corresponds to the hazard function of an individual whose risk factor vector is $x = 0$, since then $e^{\beta \cdot x} = 1$. More generally, we see that an increase of one unit in a factor x_j will multiply the hazard by a factor e^{β_j} , for which reason this is known as the “proportional hazards” regression model. It is possible to specify a parametric form for $h_0(t)$, but remarkably it turns out to be possible to estimate the terms of β without specifying the form of the h_0 , if we are willing to consider the situation immediately before a particular failure time. Again

we construct a risk set, and the chance of a particular patient failing, given the knowledge that someone in the risk set fails, provides a term in a likelihood. This is known as a “partial” likelihood since it ignores any possible information in the times between failures.

When we fit this model to the lymphoma data we find that our estimate of β for the patients with symptoms is 1.2: easier to interpret is its exponent $e^{1.2} = 3.3$, which is the proportional increase in hazard associated with presenting with symptoms. We can estimate error bounds of 1.5–7.3 around this estimate, so we can be confident that the risk of a patient who presents with symptoms will die at any stage following diagnosis is substantially higher than that of a patient who does not present with symptoms, all other factors in the model being kept constant.

A huge literature has arisen from this model, dealing with errors around estimates, different censoring patterns, tied failure times, estimating the baseline survival, and so on. Large-sample properties were rigorously established only after the method came into routine use, and have made extensive use of the theory of stochastic counting processes: see, for example, Andersen et al. (1992). These powerful mathematical tools have enabled the theory to be expanded to deal with the general analysis of sequences of events, while allowing for censoring and multiple risk factors that may depend on time.

Cox’s 1972 paper has over twenty thousand citations, and its importance to medicine is reflected in his having been awarded the 1990 Kettering Prize and Gold Medal for Cancer Research.

7 Bayesian Analysis

Bayes’s theorem is a basic result in probability theory. It states that, for two random quantities t and θ ,

$$p(\theta|t) = p(t|\theta)p(\theta)/p(t).$$

In itself this is a very simple fact, but when θ represents parameters in a model, the use of this theorem represents a different philosophy of statistical modeling. The major step in using Bayes’s theorem for inference is in considering parameters as *RANDOM VARIABLES* [III.73 §4] with probability distributions and therefore making probabilistic statements about them. For example, in the Bayesian framework one could express one’s uncertainty about a survival curve by saying that one had assessed that the probability that the mean survival time was greater than three years was 0.90. To make such an assessment, one can combine a “prior”

distribution $p(\theta)$ (a distribution representing the relative plausibility of different values of θ *before* you look at the data) with a likelihood $p(t|\theta)$ (how likely you were to observe the data t with that value of θ) and then use Bayes's theorem to provide a “posterior” distribution $p(\theta|t)$ (a distribution representing the relative plausibility of different values of θ *after* you look at the data).

Put in this way Bayesian analysis appears to be a simple application of probability theory, and for any given choice of prior distribution that is exactly what it is. But how do you choose the prior distribution? You could use evidence external to the current study, or even your own personal judgment. There is also an extensive literature on attempts to produce a toolkit of “objective” priors to use in different situations. In practice you need to specify the prior distribution in a way that is convincing to others, and this is where the subtlety arises.

As a simple example, suppose that previous studies of lymphoma had suggested that mean survival times of patients presenting with clinical symptoms probably lie between three and six years, with values of around four years being most plausible. Then it seems reasonable not to ignore such evidence when drawing conclusions for future patients, but rather to combine it with the evidence from the thirty-one patients in the current study. We could represent this external evidence by a prior distribution for θ with the form given in figure 4. When combined with the likelihood (taken from figure 2(a)), this gives rise to the posterior distribution shown. For this calculation, the functional form of the prior is assumed to be that of the *inverse-Gamma distribution*, which happens to make the mathematics of dealing with exponential likelihoods particularly straightforward, but such simplifications are not necessary if one is using simulation methods for deriving posterior distributions.

It can be seen from figure 4 that the external evidence has increased the plausibility of higher survival times. By integrating the posterior distribution above three years, we find that the posterior probability that the mean survival is greater than three years is 0.90.

Likelihoods in Bayesian models need to be fully parametric, although semi-parametric models such as the Cox model can be approximated by high-dimensional functions of nuisance parameters, which then need to be integrated out of the posterior distributions. Difficulties with evaluating such integrals held up realistic applications of Bayesian analysis for many years, but

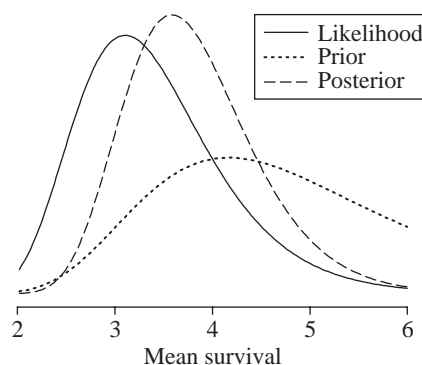


Figure 4 Prior, likelihood, and posterior distributions for mean survival time θ for patients presenting with symptoms. The posterior distribution is a formal compromise between the likelihood, which summarizes the evidence in the data alone, and the prior distribution, which summarizes external evidence that suggested longer survival times.

now developments in simulation approaches such as Markov chain Monte Carlo (MCMC) methods have led to a startling growth in practical Bayesian analyses. Mathematical work in Bayesian analysis has mainly focused on theories of objective priors, large-sample properties of posterior distributions, and dealing with hugely multivariate problems and the necessary high-dimensional integrals.

8 Discussion

The preceding sections have given some idea of the tangled conceptual issues that underlie even routine medical statistical analysis. We need to distinguish a number of different roles for mathematics in medical statistics—the following are a few examples.

Individual applications: here the use of mathematics is generally quite limited, since extensive use is made of software packages, which can fit a wide variety of models. In nonstandard problems, algebraic or numerical maximization of likelihoods may be necessary, or developing MCMC algorithms for numerical integration.

Derivation of generic methods: these can then be implemented in software. This is perhaps the most widespread mathematical work, which requires extensive use of probability theory on functions of random variables, particularly using large-sample arguments.

Proof of properties of methods: this requires the most

sophisticated mathematics, which concerns topics such as the convergence of estimators, or the behavior of Bayesian methods under different circumstances.

Medical applications continue to be a driving force in the development of new methods of statistical analysis, partly because of new sources of high-dimensional data from areas such as bioinformatics, imaging, and performance monitoring, but also because of the increasing willingness of health policy makers to use complex models: this has the consequence of focusing attention on analytic methods and the design of studies for checking, challenging, and refining such models.

Nevertheless, it may appear that rather limited mathematical tools are required in medical statistics, even for those engaged in methodological research. This is compensated for by the fascinating and continuing debate over the underlying philosophy of even the most common statistical tools, and the consequent variety of approaches to apparently simple problems. Much of this debate is hidden from the routine user. Regarding the appropriate role of mathematical theory in statistics, we can do no better than quote David Cox in his 1981 Presidential Address to the Royal Statistical Society (Cox 1981):

Lord Rayleigh defined applied mathematics as being concerned with quantitative investigation of the real world “neither seeking nor evading mathematical difficulties.” This describes rather precisely the delicate relation that ideally should hold between mathematics and statistics. Much fine work in statistics involves minimal mathematics; some bad work in statistics gets by because of its apparent mathematical content. Yet it would be harmful for the development of the subject for there to be widespread an anti-mathematical attitude, a fear of powerful mathematics appropriately deployed.

Further Reading

- Andersen, P. K., O. Borgan, R. Gill, and N. Keiding. 1992. *Statistical Models Based on Counting Processes*. New York: Springer.
- Cox, D. R. 1972. Theory and general principle in statistics. *Journal of the Royal Statistical Society A* 144:289–97.
- . 1981. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B* 34: 187–220.
- Kaplan, E. L., and P. Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53:457–81.

- Matthews, D. E., and V. T. Farewell. 1985. *Using and Understanding Medical Statistics*. Basel: Karger.

VII.12 Analysis, Mathematical and Philosophical

John P. Burgess

1 The Analytic Tradition in Philosophy

Philosophical problems are never solved for the same reason that treasonous conspiracies never succeed: as successful conspiracies are never called “treason,” so solved problems are no longer called “philosophy.” Philosophy, which once included almost every subject in the university (every subject in which the highest degree is Ph.D.), has thus been shrunk by success. The greatest shrinkage occurred during the seventeenth and eighteenth centuries, when natural philosophy became natural science. Philosophers of the period, all intensely interested in the emergence of the new science, differed over issues of scientific method. Philosophy had always been understood to differ from, for instance, theology, by restricting itself to the methods of reasoned argument and the evidence of experience, without appeal to authority, tradition, revelation, or faith. But philosophers of the era of the scientific revolution disagreed about the comparative importance of reason and experience.

In introductory histories, philosophers are accordingly divided into the rationalists, or the party of reason, and the empiricists, or the party of experience. The former, mainly from Continental Europe, were dominant in the seventeenth century, while the latter, mainly from the British Isles, predominated in the eighteenth. The rationalists, who included the mathematicians DESCARTES [VI.11] and LEIBNIZ [VI.15], were impressed by the apparent ability of pure thought—logical deduction from self-evident postulates—to achieve, as it seemed to do in geometry, substantive results with worldly applications; and they were tempted to adopt similar methods in other areas. Spinoza even wrote his *Ethics* in the style of EUCLID’s [VI.2] *Elements*, a world-historic peak of the influence of mathematics on philosophy. The empiricists, who included that acute critic of the calculus, Berkeley, recognized that in physics one cannot proceed as the rationalists wished to. The principles of physics are not self-evident, but must be conjectured from and tested against systematic observation and controlled experiment. What puzzled leading

empiricists such as Locke and Hume was how pure thought was able to succeed in *any* area, as it seemed to in geometry. Thus, while for rationalists mathematics was a source of *methods*, for empiricists it was the source of a *problem*.

An influential formulation of that problem was offered by Kant, whose system attempted a synthesis of rationalism with empiricism. On the one hand, Kant claimed, geometry and arithmetic are a priori rather than a posteriori, by which he meant that they are knowable in advance of experience rather than dependent on it. On the other hand, they are synthetic rather than analytic, which is to say that they are more than mere logical consequences of the definitions of concepts, statements whose denials would amount to contradictions in terms. Philosophy of mathematics, today a smallish specialty within philosophy of science, itself a smallish specialty within epistemology or the theory of knowledge, played a much more important role for Kant, who in his own summary of his system gave pride of place to the question, "How is pure mathematics possible?" as the first case of the question, "How is synthetic a priori knowledge possible?" Kant's proposed solution was based on the insight that our knowledge must be shaped as much by the nature of ourselves, the knowers, as by that of what is known. Kant concluded that space, the subject matter of geometry, and time, according to him the ultimate subject matter of arithmetic, were not features of things as they are in themselves, but rather of things as we must perceive and experience them, given the nature of our sensibility. Synthetic a priori knowledge is ultimately *self-knowledge*, knowledge of the forms that *we* supply, and into which reality independent of us pours content. This distinction between *phenomena*, or things as we experience them, and *noumena*, things beyond our experience, about which we can wonder but never know, was central to Kant's entire system, his ethics as much as his metaphysics.

Such is the history of early modern philosophy, painted in quick strokes with a broad brush. After Kant, the story no longer has as clear a plotline. System building continued for another generation, down to Hegel. But eventually, and inevitably, his system collapsed under its own weight, and in the ensuing reaction philosophers went off in all directions. Outside academia, striking figures sporadically appeared on the borders of philosophy and literature, notably Nietzsche. Meanwhile, academic philosophy, rather like Victorian architecture, experienced a number of

revivals, of which the Kantian was the most prominent. But even as neo-Kantianism prevailed in the schools, the Kantian conception of mathematics was under attack. First, though the development of consistent non-Euclidean geometries in itself only confirms Kant's claim that geometry is synthetic, those who developed alternatives to Euclid were quickly led to question whether Euclidean geometry is really a priori, as Kant had claimed. GAUSS [VI.26] had already concluded that geometry is a posteriori, or, as he put it, of the same status as mechanics, and RIEMANN [VI.49] argued at greater length that an examination of the hypotheses that lie at the foundation of geometry must lead us into the domain of the neighboring science of physics. Second, while few doubted Kant's claim that arithmetic is a priori, a challenge arose to the claim that it is synthetic in the work of GOTTLIEB FREGE [VI.56] and (slightly later, but largely independently) BERTRAND RUSSELL [VI.71], who both attempted a derivation of arithmetic from logic along with an appropriate definition of number.

Frege's work long remained less well-known than it deserved to be, despite the publicity given it by Russell once he became aware of it himself. As a result, Frege, though very influential at present, is more a precursor of the tradition in philosophy within which he stands than a founder, the founders being rather Russell and his contemporary and colleague G. E. Moore. That pair began by rebelling against the philosophy of their teachers, a late nineteenth-century aberration called absolute idealism, a kind of Hegel revival; but it soon became apparent that the rebels were aiming at more than just a return to the traditional empiricism of British philosophy from Bacon to Mill. Meanwhile, Edmund Husserl was developing the first form of what was to become the great rival to the Russell-Moore tradition in twentieth-century philosophy. Like Frege, Husserl had begun his career with work in the philosophy of arithmetic, work of which Frege himself had taken notice, and no one in the early twentieth century expected that Husserl's and Frege's heirs would, within a generation, split into two noncommunicating lines of descent.

The two lines of development or traditions are oddly named, with a stylistic label, "analytic," for one, and a geographical label, "Continental," for the other. This odd labeling reflects the historical fact that the principal representatives of the analytic style in Continental Europe (Ludwig Wittgenstein, Rudolf Carnap, and others) were forced to go into exile in the English-

speaking world in the 1930s, as a result of the process generally known as the Nazification (but celebrated by Husserl's estranged student Martin Heidegger as the "self-affirmation") of the German university. This physical separation—more than Heidegger's break with his teacher, hostility toward science, rebarbative prose style, or loathsome politics—created a split that no one could have anticipated twenty years earlier.

With the years the gap has widened, as later writers on each side tend to read and cite only predecessors on that side. Indeed, the divide has extended backwards in time. For while Borges has said that in literature great writers create their own predecessors, in philosophy even not-so-great writers can do so, and the two twentieth-century traditions came to see different nineteenth-century figures as leading up to themselves, thus extending the division between them right back to the death of Kant (with Hegel rather than Heidegger being identified as the first distinctively Continental philosopher). The gap between the reading lists of students in the two traditions has become so large that nowadays for a student trained in one to take up the other is virtually to switch disciplines.

The word "tradition," rather than "school" or "movement," is used advisedly, for each tradition has contained several movements, as well as individuals who defy classification by school. It would be a serious mistake to suppose that there is any doctrine or method on either side of the analytic/Continental divide that all philosophers on that side uphold. In particular, analytic philosophy should not be confused with logical positivism, a Viennese-American school defunct for more than half a century, nor should Continental philosophy be confused with existentialism, a literary-philosophical movement out of fashion in Paris for nearly as long. Logical positivism and existentialism were indeed varieties of analytic and Continental philosophy respectively, and perhaps the most prominent varieties half a century or so ago; but each was even then far from being the *only* variety. In assessing the influence of mathematics on philosophy in the twentieth century, one must take into account divisions within each tradition as much as the division between the two traditions.

It may be true that since the early work of Husserl there has been comparatively little contact between mathematics and philosophy on the Continental side, though the label "structuralist" is certainly broad enough to take in both the mathematics of BOURBAKI [VI.96] and the various anthropological and linguistic

doctrines that became influential in France after the eclipse of existentialism; but it is also true that the direct influence of mathematical ways of thinking on many individuals and groups within the analytic tradition has been negligible. Thus, just as there are distinguishable German and French subtraditions within the Continental tradition, so within the analytic tradition one may distinguish a more technically oriented subtradition, including Frege (who was himself a professor of mathematics), Russell (who as an undergraduate had concentrated on mathematics before turning to philosophy), and the logical positivists (who had mostly been trained as theoretical physicists), from a nontechnical or antitechnical subtradition, including Moore, Wittgenstein, the so-called ordinary-language school of mid-century Oxford, and others. (Wittgenstein even went so far as to claim that mathematicians always make bad philosophers, a sweeping judgment condemning many right back to Thales and PYTHAGORAS [VI.1], though the immediate target was Russell.) However, there has been very much more communication and influence back and forth between the two subtraditions within each tradition than between the two traditions.

Even among the more technical analytic philosophers the influence of mathematics after the period of the founders has been occasional and sporadic, and has come mostly from areas such as mathematical logic, computability theory, probability and statistics, game theory, and mathematical economics (as in the work of the philosopher-economist Amartya Sen), which are rather far from the core of pure mathematics as mathematicians see it. Thus it is hard to imagine the solution to any of the Millennium Prize Problems (except perhaps the P vs NP problem, the one question coming from theoretical computer science rather than core mathematics) having measurable impact even among the most susceptible analytic philosophers. In contrast to this limited *direct* influence, the *indirect* influence of mathematics, resulting from its effect on the thought of the early figures Frege and Russell, has been overwhelming even among the less technically oriented analytic philosophers. The branches of mathematics that influenced Frege and Russell were geometry and algebra and, above all, the third great branch of core mathematics, "analysis," in the mathematical rather than the philosophical sense, the branch beginning with differential and integral calculus. (Frege and Russell were not *influenced* by mathematical logic: rather, they *created* it, and mathematical analysis was a key influence on its creation.)

2 Mathematical Analysis and Frege's New Logic

Let us turn, then, to consider the state of mathematical analysis in the days of Frege and Russell, beginning our account with a quick look back at the situation ca. 1800. As rich as its results were, and as powerful its applications, mathematics at the beginning of the nineteenth century was concerned with but a few structures: the natural, rational, real, and complex number systems; and the Euclidean and projective spaces of dimensions one, two, and three. All that changed quickly when the work of GAUSS, HAMILTON [VI.37], and others introduced the first non-Euclidean spaces and first noncommutative algebras, after which a proliferation of new mathematical structures rapidly ensued. This *generalizing* tendency went hand in hand with a *rigorizing* tendency, since the proliferation of novelties persuaded mathematicians that they needed to adhere more strictly than had become customary to the ancient ideal of rigor, according to which all new results in mathematics are to be logically deduced from previous results, and ultimately from a list of explicit axioms. For without rigor, intuitions derived from familiarity with more traditional structures might easily be unconsciously transferred to new situations where they are no longer appropriate.

Generalization and rigorization went hand in hand not only in geometry and algebra, but also in mathematical analysis. Generalization in mathematical analysis took place in two directions. The eighteenth-century notion of "function" had been that of an operation applying to one or more real numbers as inputs or "arguments" and yielding a real number as output or "value," according to a certain formula, such as $f(x) = \sin x + \cos x$ or $f(x, y) = x^2 + y^2$. On the one hand, nineteenth-century mathematicians generalized by dropping the requirement of an explicit formula. On the other hand, Cauchy, Riemann, and others extended the notion to allow as arguments not only real numbers but also complex numbers, that is, numbers of the form $a + bi$, where a and b are real numbers and i is the "imaginary" square root of -1 .

Rigorization in mathematical analysis also took place on two levels. First, for each theorem it had to be clearly stated just what special properties were being assumed for the functions to which the result was supposed to apply, since special properties such as definability by a formula (or continuity or differentiability) were no longer being built into the highly general notion

of function itself; moreover, the relevant properties themselves had to be clearly defined (leading to the so-called WEIERSTRASS [VI.44] epsilon-delta definitions of such concepts as "continuity" and "differentiability" in freshman calculus), since, as POINCARÉ [VI.61] remarked, until one has rigor in one's definitions one cannot have rigor in one's theorems. Second, the properties assumed for the numbers to which the functions apply had also to be clarified and stated explicitly as axioms, with the properties of complex numbers being derived by logical definition and deduction from properties of real numbers (by HAMILTON), which themselves in turn were derived from properties of rational numbers (by DEDEKIND [VI.50] and CANTOR [VI.54]), which themselves in turn were derived from properties of the system of natural numbers 0, 1, 2, and so on.

Here Frege wished to press still further, and to do what Kant had said could not be done, and derive the properties of the natural numbers themselves from pure logic. For this purpose he needed to become more self-conscious about logic than even the most rigorist mathematicians: he needed not merely to adhere implicitly to the rules and standards of logical definition and deduction, but also to analyze explicitly those very rules and standards themselves. Such self-conscious analysis of definition and deduction was a topic that had, since antiquity, traditionally belonged to philosophy rather than mathematics. Frege needed to carry out a revolution in this philosophical subject, one that would bring it much closer to mathematics, and would bring progress to a field that Kant had described as having advanced not a step beyond the state in which it was left by its founder, Aristotle. (The description is slightly exaggerated, but essentially correct, in that each step forward in the two millennia after Aristotle had been followed by a step back.) It was Frege's new logic, detached from its original role as part of a special project in foundations of arithmetic and applied to quite diverse subject matters, that was to become the single most important general instrument for philosophical analysis in the twentieth century. Indeed, to a large degree philosophical analysis simply *is* the logical analysis of philosophical rather than mathematical notions, carried out with the aid of Frege's broad new logic, or still broader extensions of it introduced by his successors. It was by the creation of this general instrument of a new logic, rather than the specialized application he made of it to the philosophy of mathematics, that Frege became the grandfather of analytic philosophy. And the novelty in Frege's

logic was directly inspired by novel developments in mathematical analysis, as he himself emphasized.

In an article entitled "Function and concept," Frege describes the broadening of the notion of function as follows (in the translation by Peter Geach and Max Black):

Now how has the reference of the word "function" been extended by the progress of science? We can distinguish two directions in which this has happened. In the first place, the field of mathematical operations that serve for constructing functions has been extended. Besides addition, multiplication, exponentiation, and their converses, the various means of transition to the limit have been introduced—to be sure, without people's being always clearly aware that they were thus adopting something essentially new. People have even gone further still, and have actually been obliged to resort to ordinary language, because the symbolic language of Analysis failed, e.g., when they were speaking of a function whose value is 1 for rational and 0 for irrational arguments. [This is a famous example of DIRICHLET [VI.36].] Secondly, the field of possible arguments and values for functions has been extended by the admission of complex numbers. In conjunction with this, the sense of the expressions "sum," "product," etc. had to be defined more widely.

Frege adds at the end, "In both directions I go still further." For it was the broadening of the notion of function by mathematicians that provided Frege with the clue he needed to develop a logic broader than Aristotle's.

Before one can appreciate the advance represented by Frege's logic, one must understand something of Aristotle's. Though it is a pretty poor achievement if it is considered as the best the human race could do in this area in a couple of thousand years, it is a brilliant one when considered as the work of a single individual in the course of a career devoted to many other projects. For Aristotle created from nothing the science of logic, whose aim is to distinguish valid from invalid inferences of conclusions from premises. Here an inference is valid if its form alone, regardless of the material truth or falsehood of premises and conclusions, guarantees that *if* the premises are true, *then* the conclusion is true. Equivalently, the inference is valid if in all inferences of the same form in which the premises are true, the conclusion is true. Thus, to adapt an example of Lewis Carroll, the inference from "I believe whatever I say" to "I say whatever I believe" is *not* valid, because there are inferences of identical form in which the premise is true and the conclusion false, such as the

inference from "I see whatever I eat" to "I eat whatever I see."

The scope of Aristotle's logic is limited by the limited range of forms of potential premises and conclusions he recognizes. In fact, he recognized only four: the *universal affirmative* "All A's are B's," the *universal negative* "No A's are B's," the *particular affirmative* "Some A's are B's," and the *particular negative* "Some A's are not B's" or "Not all A's are B's." The premise "I believe whatever I say" amounts to "All things that I say are things that I believe," and hence is a universal affirmative. The invalidity of the inference in the Lewis Carroll example exemplifies the invalidity of the inference from "All A's are B's" to "All B's are A's." The validity of the inference from the two premises "All Greeks are human beings" and "All human beings are mortal" to the conclusion "All Greeks are mortal" exemplifies the validity of the inference from "All A's are B's" and "All B's are C's" to "All A's are C's," traditionally called the "syllogism in *Barbara*," for reasons that need not concern us here. Aristotle's logic was in part inspired by the practice of deduction in philosophical debate ("dialectic") and in part by the practice of deduction in mathematical theorem-proving ("demonstration"), and he offers in his *Posterior Analytics* an account of a deductive science that is presumed to be based on the practice of the contemporary geometer Eudoxus, in the same sense and to the same degree in which his account in the *Poetics* of tragedy is based on the practice of the contemporary playwright Euripides. But, in fact, Aristotle's logic is inadequate for the analysis of mathematicians' actual arguments, because he makes no provision for forms of argument involving *relations*. He cannot, for instance, analyze properly the valid argument from "All squares are rectangles" to "Anyone who draws a square draws a rectangle," because he has no way of representing adequately the form of the conclusion.

By contrast, if you open any present-day introductory logic text, you will find instructions on how to represent symbolically the forms of arguments involving relations. The example just given would appear textbook-style as follows:

$$\begin{aligned} &\forall x(\text{Square}(x) \rightarrow \text{Rectangle}(x)) \\ \therefore &\forall y(\exists x(\text{Square}(x) \ \& \ \text{Draws}(y, x)) \rightarrow \\ &\quad \exists x(\text{Rectangle}(x) \ \& \ \text{Draws}(y, x))). \end{aligned}$$

In words this would amount to the following. For every x , if x is a square, then x is a rectangle. Therefore, for every y , if there is an x such that x is a square

and y draws x , then there exists an x such that x is a rectangle and y draws x . (Thus “ \rightarrow ” means “if . . . , then . . . ,” “ \forall ” means “for every,” and “ \exists ” means “there is.”) This style of logical analysis is the invention of Frege.

Underlying it is a notion of a “concept” as a special kind of function, a function that (generalizing the mathematical notion in one direction) need not be given by any kind of *mathematical* description, and that (generalizing the mathematical notion in another direction) need not have as arguments any kind of *numbers*. A concept for Frege is a function whose argument or arguments may be any objects at all, and whose values are Truth and Falsehood. Thus, the concept Wise applied to the argument Socrates produces the value Truth, since Socrates is wise (at least to the extent of recognizing that he lacked perfect wisdom), while the concept Immortal applied to Socrates produces Falsehood, since Socrates was not immortal but died of drinking hemlock. Frege is able to handle relations *because he follows the mathematical analysts who allowed functions of two or more arguments*. Thus the two-argument concept or relation Taught applied to Socrates and Plato, in that order, produces Truth, since Socrates taught Plato, while applied to Plato and Socrates, in *that* order, produces Falsehood, since Plato did not teach Socrates. Aristotle’s simple “All A’s are B’s” becomes, for Frege, the more complex “For all objects x , if $A(x)$, then $B(x)$.” At the price of such extra complexity, he is able to logically analyze arguments turning on relations, as Aristotle was not.

Aristotle analyzed the concept Human Being in terms of the concepts Animal and Rational in the sense of “language-using.” In present-day textbook notation (writing “ \leftrightarrow ” for “if and only if”), this would be

$$\text{Human}(x) \leftrightarrow \text{Animal}(x) \ \& \ \text{Rational}(x).$$

But Aristotle, with no theory of relations, was unable to analyze the notion of Mother (respectively, Father) in terms of Female (respectively, Male) and Parent. For Frege, Mother is analyzed as follows:

$$\text{Mother}(x) \leftrightarrow \text{Female}(x) \ \& \ \exists y \text{Parent}(x, y).$$

A mother is a female who is someone’s parent, and analogously for a father. Frege was even able to analyze the concept Ancestor in terms of the concept Parent, though this analysis is beyond the scope of the present sketch. Later philosophical analysis would have been unthinkable without Frege’s broadening of logical analysis beyond Aristotle’s, and Frege rightly saw

his broadening of logical analysis as a direct extrapolation from the nineteenth-century mathematical analysts’ broadening of the notion of function they had inherited from their eighteenth-century predecessors.

3 Mathematical Analysis and Russell’s Theory of Descriptions

Like Frege, Russell found in mathematics both a source of problems and a source of methods. For the purposes of a specialized investigation of problems in the philosophy of mathematics, he created an instrument, his theory of descriptions, and a more general method, that of contextual definition, which his successors took up and applied to many other problem areas. Indeed, it was not merely Russell’s successors who applied these ideas to areas outside philosophy of mathematics, since Russell himself did so in his first publications on the subject. Thus it is not apparent from Russell’s still widely read “On denoting,” published in 1905 and even today a key item on the syllabus of students of analytic philosophy, that the theory of descriptions originated in the course of studies in foundations and philosophy of mathematics. Rather, this is a fact mentioned in Russell’s autobiographical writings and known to historians of twentieth-century philosophy. The degree to which the method of contextual definition, which the theory of descriptions exemplifies, was inspired by the nineteenth-century rigorization of analysis is perhaps not sufficiently appreciated even by such specialists.

A principal puzzle Russell addresses in “On denoting” is that of so-called negative existentials, such as “The king of France does not exist.” In superficial grammatical form this statement resembles “The queen of England does not agree,” and to that extent it appears to involve picking out an object (in this case, a person), and then attributing a property to him (or her, as the case may be). Thus it seems that in order to say that someone or something does not exist, one must assume that in some sense there is such a person or thing, to whom or which the property of nonexistence may be ascribed. Russell cites Alexius Meinong (a student of Husserl’s teacher Franz Brentano) as a philosopher committed to such a view. For Meinong had a theory of “objects beyond being and nonbeing,” exemplified by The Golden Mountain and The Round Square. But as Scott Soames reveals, in his *Philosophical Analysis in the Twentieth Century*, volume I: *The Dawn of Analysis*, Russell himself had briefly held a similar view in

the first days of his and Moore's joint rebellion against absolute idealism. It was through the development of his theory of descriptions that Russell was able to free himself from anything like commitment to Meinongian "objects."

According to that theory, to say that *a* Golden Mountain exists is to say that there is something that is both golden and a mountain: $\exists x(\text{Golden}(x) \ \& \ \text{Mountain}(x))$. To say that *the* Golden Mountain exists is to say that there is one thing that is both golden and a mountain and no other such thing:

$$\begin{aligned} &\exists x(\text{Golden}(x) \ \& \ \text{Mountain}(x)) \\ &\ \& \ \sim \exists y(\text{Golden}(y) \ \& \ \text{Mountain}(y) \ \& \ y \neq x). \end{aligned}$$

(Here " \sim " represents "it is not the case that.") This is logically equivalent to saying there is something such that a thing is both golden and a mountain if and only if it is identical with that thing:

$$\exists x \forall y (\text{Golden}(y) \ \& \ \text{Mountain}(y) \leftrightarrow y = x).$$

To say that the Golden Mountain does *not* exist is simply to deny this:

$$\sim \exists x \forall y (\text{Golden}(y) \ \& \ \text{Mountain}(y) \leftrightarrow y = x).$$

To say that the king of France is bald is, similarly, to say that there is something such that a thing is king of France if and only if it is identical with that thing, and that thing is bald:

$$\exists x (\forall y (\text{King-of-France}(y) \leftrightarrow y = x) \ \& \ \text{Bald}(x)).$$

This is not the place to go into the subtleties of Russell's theory, whose main point should be clear from these few examples: when the logical form is properly analyzed, using the new logic, the phrase "the Golden Mountain" or "the present king of France" disappears. With it vanishes any appearance that we must acknowledge such an "object" as the Golden Mountain or king of France even in order to deny that any such object exists. The examples illustrate in miniature two lessons: first, that the logical form of a statement may differ significantly from its grammatical form, and that recognition of this difference may be the key to solving or dissolving a philosophical problem; second, that the correct logical analysis of a word or phrase may involve an explanation not of what *that word or phrase taken by itself* means, but rather of what *whole sentences containing the word or phrase* mean. Such an explanation is what is meant by a *contextual* definition: a definition that does not provide an analysis of the word or phrase standing alone, but rather provides an analysis of contexts in which it appears.

Russell's distinction between grammatical and logical form, and his claim that the former may be systematically misleading, was to prove immensely influential, even among nontechnically oriented philosophers, such as the Oxford ordinary-language school, who saw no need to use special symbols to represent logical forms, and objected to details of Russell's specific application of the distinction in his theory of descriptions. But Russell's notion of contextual definition is one implicit already in the practice of Weierstrass and other leaders of the nineteenth-century rigorization of analysis, and familiar to Russell from his undergraduate mathematical studies, so that even the antitechnical ordinary-language school of philosophical analysts are being influenced at one remove (and, so to speak, in spite of themselves) by mathematical analysis.

Contextual definition was the tool the rigorizers used to dispel the mysteries surrounding the notions of infinitesimals and infinities in the calculus. The followers of Leibniz had, for instance, written $df(x)/dx$ for the derivative of a function $f(x)$, wherein dx was supposed to represent an "infinitesimal" change in the argument, and $df(x)$ a corresponding "infinitesimal" change $f(x + dx) - f(x)$ in the value when the argument changes from x to $x + dx$. (Leibniz claimed that this was all just a figure of speech, but his followers seem to have taken it literally.) These infinitesimals could be treated as nonzero in some circumstances—in particular, one could divide by them, as one cannot divide by zero—and yet treated as zero and neglected in other circumstances. Thus the derivative of the function $f(x) = x^2$ was computed as follows:

$$\begin{aligned} \frac{df(x)}{dx} &= \frac{f(x + dx) - f(x)}{dx} = \frac{(x + dx)^2 - x^2}{dx} \\ &= \frac{2x dx + (dx)^2}{dx} = 2x + dx = 2x. \end{aligned}$$

Here dx is treated as nonzero at the next-to-last step, and zero at the last step—the kind of procedure that outraged critics like Berkeley. In the course of the nineteenth-century rigorization, the infinitesimals were banished: what was provided was not a direct explanation of the meaning of $df(x)$ or dx , taken separately, but rather an explanation of the meaning of contexts containing such expressions, taken as wholes. The apparent form of $df(x)/dx$ as a quotient of infinitesimals $df(x)$ and dx was explained away, the true form being $(d/dx)f(x)$, indicating the application of an operation of differentiation d/dx applied to a function $f(x)$.

Similarly, such an expression as $\lim_{x \rightarrow 0} 1/x = \infty$, or “the limit of $1/x$ as x goes to zero is infinity,” was explained *as a whole*, without requiring any explanation of “ ∞ ” or “infinity” taken separately. The details, which now appear in any freshman calculus textbook, need not detain us. What is important historically is that the notion of contextual definition employed in Russell’s theory of descriptions was an idea that would have been familiar to him as a student of mathematics. To acknowledge this is, needless to say, not to deny that there is a certain genius involved in extracting such an idea from its original context of mathematical analysis and employing it to resolve philosophical puzzles. To acknowledge the germs of Russell’s ideas in ideas of Weierstrass is merely to indicate more precisely what *kind* of genius Russell, like Frege before him, was bringing to bear on philosophical issues: a kind of philosophical genius *informed by knowledge of mathematics*.

4 Philosophical Analysis and Analytic Philosophy

Anyone who acquires a new tool is in some danger of behaving like the proverbial man with a hammer to whom everything seems to be a nail. There is no denying that some of the first people to apply the new methods of Frege and Russell were overenthusiastic about what such methods could accomplish. Russell himself, having established to his own satisfaction that mathematics could be reduced to pure logic once one had a sufficiently rich and powerful logic, went on to conclude that every science apart from mathematics could be reduced to logical compounds of statements about immediate sensory impressions—“sense data” as they were called. The logical positivists reached a similar conclusion, and were ready to ban any statement that did not admit such a reduction, from the assertions of Hegelian or absolute idealist metaphysicians on, as a “pseudo-statement,” or mere nonsense.

Conscientious attempts to work out just *how* science, even the parts concerned with theoretical entities not directly observable (such as quarks and black holes in the science of today), could be reduced logically to statements about sense data, or at least to statements about everyday observable objects (such as meter readings), failed. Hence the positivists were forced to acknowledge that their program could not succeed, and (since they did not wish to dismiss large parts of modern science as mere pseudo-statements)

that their standards of meaningfulness were too rigid. But as Soames emphasizes, this very acknowledgment of failure was a kind of success, because few if any philosophical schools before the positivists had even stated their aims with sufficient clarity to make it possible to see that they were unachievable. The new logical resources provided by Frege and Russell had *both* tempted the positivists to conjecture more than they could prove *and* made it clear to them that proof of their conjecture was impossible.

With experience the scope and limits of the new methods gradually came to be better understood. Russell’s theory of descriptions had been hailed by his student F. P. Ramsey as “a paradigm of philosophical analysis,” which indeed it is. But it came to be appreciated that the kind of application Russell made to the issue of negative existentials, where a philosophical problem was completely dissolved by philosophical analysis, would seldom be possible. Analysis, in general, is only a preliminary, a process that makes it clearer what the real problems are, and not a panacea, exposing all apparent problems as mere pseudo-problems.

As analytic philosophy has developed, enthusiasm has been replaced by dedication: recognition of the limitations of Frege’s and Russell’s methods has led not to the abandonment of the goal of clarity, which was the underlying motive of the great pioneering figures, but rather to firmer adherence to it. Today, when one can read large tracts of philosophy in the analytic tradition without encountering a single explicit analysis, let alone one expressed in special logical symbolism, one still finds almost everywhere a clarity of prose style that instantly distinguishes writing in this tradition from the writings of Continental philosophers (to say nothing of the Continentalizing philosophastersto be found in certain humanities departments in universities in the English-speaking world). This clarity—found, to be sure, already in the mathematician-philosopher Descartes, the first truly modern philosopher, but lost in many of his successors—is the ultimate influence and legacy which the pioneers of analytic philosophy transmitted from mathematics to their philosophical heirs.

Further Reading

I recommend *Philosophical Analysis in the Twentieth Century* (Princeton, NJ: Princeton University Press, 2003) by Scott Soames for those wishing to read more

about this subject. Each of the two volumes of this work contains substantial lists of primary and secondary sources at the end of each of its several parts.

VII.13 Mathematics and Music

Catherine Nolan

1 Introduction and Historical Overview

Music is the pleasure the human mind experiences from counting without being aware that it is counting.

This intriguing remark of LEIBNIZ [VI.15], from a 1712 letter to fellow mathematician Christian GOLDBACH [VI.17], suggests a serious connection between mathematics and music, two subjects—one a science, the other an art—that may at first seem very different from each other. Leibniz was perhaps thinking of the long-standing historical and intellectual association of the two disciplines that date back to the time of PYTHAGORAS [VI.1], when the subject of music was part of an elaborate classification scheme of knowledge in the mathematical sciences. This scheme became known in the Middle Ages as the *quadrivium*, and consisted of the four disciplines of arithmetic, music (harmonics), geometry, and astronomy. In the Pythagorean worldview, these subjects were interlinked, since in one way or another they were all concerned with simple ratios. Music was merely the aural manifestation of a more universal harmony, which was likewise expressed by relationships between numbers, geometrical magnitudes, or the motions of celestial bodies. Harmonic consonance of musical intervals resulted from simple ratios of the first four natural numbers, 1:1 (the unison), 2:1 (the octave), 3:2 (the perfect fifth), and 4:3 (the perfect fourth), and was demonstrated empirically by the ratios of lengths of vibrating strings on the ancient instrument the monochord.¹ Beginning with the Scientific Revolution of the seventeenth century, theories of tuning and temperament of musical intervals required more advanced mathematical ideas as well, such as logarithms and decimal expansions.

Musical composition has been inspired by mathematical techniques throughout its history, although mathematically inspired compositional techniques are associated mainly with music of the twentieth, and

now twenty-first, centuries. A striking early example appears in the section on melody in a monumental treatise on music, entitled *Harmonie universelle* (1636–37), by the mathematician Marin Mersenne. Mersenne applied simple (from today's perspective) combinatorial techniques to the distribution and organization of notes in melodies. For example, he calculated the number of different arrangements or permutations of n notes, for each n between 1 and 22 (twenty-two notes delimiting the range of three octaves). The answer is of course $n!$, but in his zeal to illustrate this, he notated on musical staves all 720 ($6!$) permutations of the six notes of the minor hexachord (A, B, C, D, E, F), occupying a full twelve pages of *Harmonie universelle*. He went on to explore more complicated problems such as determining the number of melodies of a certain number of notes selected from a larger number, or determining the number of arrangements of finite collections of notes containing certain numbers of repetitions of one or more notes. He illustrated some of his findings with combinations of letters as well as musical notation, thereby showing that the music was incidental to the problems, which were in essence purely combinatorial. Such exercises, while seemingly of little practical or aesthetic value, at least demonstrated the great musical diversity that was in principle available with only a limited set of resources.

The polymath Mersenne was a composer and practicing musician as well as a mathematician, and his fascination with applying a relatively new mathematical technique to music composition showed a level of interest in abstract connections between mathematics and music that is shared by many music theorists, and to a lesser degree by performing musicians and nonspecialist music enthusiasts. The patterns of music, in particular pitch and rhythm, lend themselves well to mathematical description, and some of them are amenable to algebraic reasoning. In particular, the system of twelve equal-tempered notes is naturally modeled using MODULAR ARITHMETIC [III.60], and this, together with combinatorial arguments, was used in the music theory of the twentieth century. In this article we survey the association of mathematics and music from its concrete representation in sound itself, through its manifestation in the working materials of composers, and finally to its explanatory power in abstract music theory.

2 Tuning and Temperament

The most obvious relationships between mathematics and music appear in acoustics, the science of musi-

1. The monochord was an instrument designed for demonstration, not artistic, purposes. It consisted of a single string stretched between two fixed bridges. A movable bridge between the fixed bridges was used to adjust the length of the string as it was plucked to produce sound, thereby altering the pitch of the sound.

cal sound, and particularly in the analysis of the intervals between pairs of pitches. With the development of polyphonic music in the Renaissance period, the Pythagorean conception of consonance based on the simple ratios of the integers from 1 to 4 eventually came into conflict with musical practice. The acoustically pure perfect consonances of Pythagorean tuning were well-suited for medieval parallel organum,² but in the fifteenth and sixteenth centuries use was increasingly made of the so-called *imperfect consonances*, that is, major and minor thirds and their octave inversions, minor and major sixths. In Pythagorean tuning, intervals are derived by successions of perfect fifths, so the corresponding frequency ratios are powers of $\frac{3}{2}$. In conventional Western music, twelve perfect fifths in succession, C–G–D–A–E–B–F $^\sharp$ –C $^\sharp$ –G $^\sharp$ –D $^\sharp$ –A $^\sharp$ –E $^\sharp$ –B $^\sharp$, are supposed to equal seven octaves (C = B $^\sharp$), but this does not work in Pythagorean tuning, since $(\frac{3}{2})^{12}$ does not equal 2^7 . Indeed, a succession of Pythagorean perfect fifths will never result in a whole number of octaves. As it happens, twelve Pythagorean perfect fifths give an interval slightly larger than seven octaves. The difference is a small interval known as the *Pythagorean comma*, which corresponds to a ratio of $(\frac{3}{2})^{12}/2^7$, which is about 1.013643.

Pythagorean tuning was originally conceived in terms of successive single pitches. The problems associated with it start to arise when pitches sound simultaneously. While Pythagorean fifths between simultaneous pitches sound pleasing with their simple 3:2 ratios, Pythagorean thirds and sixths have much more complex ratios that sound harsh to Western ears. These came to be replaced by the simple ratios of *just intonation*, which are ratios of quite small whole numbers. These ratios were considered “natural” because they reflect the ratios of the natural overtone series.³ The Pythagorean major third, which has the relatively complex ratio of $(\frac{3}{2})^4/2^2$, or $\frac{81}{64}$, was replaced by the slightly smaller major third of just intonation, which has the much simpler ratio 5:4. The difference between these two intervals is known as the *syntonic comma*, which corresponds to the ratio 81:80, or 1.0125. Likewise,

2. *Organum* is the earliest form of musical polyphony, and involved adding a voice (or voices) to an existing plainchant melody (*cantus firmus*). In its original form, the added voice proceeded in parallel motion to the plainchant melody at the interval of a perfect fourth or fifth.

3. The partials of the overtone series are multiples of the frequency of the fundamental pitch, and the first six partials generate the intervals of the major triad. For instance, the first six partials of the overtone series of a fundamental pitch C are C (1:1), C (2:1), G (3:1), C (4:1), E (5:1), G (6:1).

Notes	C	D	E	F	G	A	B	C
Intervals (ratios)	$\frac{9}{8}$	$\frac{10}{9}$	$\frac{16}{15}$	$\frac{9}{8}$	$\frac{10}{9}$	$\frac{9}{8}$	$\frac{16}{15}$	

Figure 1 Successive intervals in a major scale tuned in just intonation.

the Pythagorean minor third has ratio 32:27, and so is slightly smaller than the minor third of just intonation, which has ratio 6:5. The difference is again a syntonic comma. The Pythagorean major and minor sixths, the octave inversions of the thirds, also differ from their just counterparts by a syntonic comma.

Suppose that you want to build a C-major scale in just intonation. You can do it as follows. Start with C and define each other note by the ratio of its frequency to that of C. The subdominant and dominant, that is, F and G, have ratios 4:3 and 3:2, respectively. From these three notes one can build major triads in the ratios 4:5:6. So E, for instance, which belongs to the major triad that starts with C, has ratio 5:4. Similarly, A has ratio 5:3, since it is in a ratio 5:4 with F. With this kind of calculation, one ends up with the scale shown in figure 1, where the fractions now represent the frequency ratios between successive notes. The smaller whole tone (10:9) between notes D and E creates intonation problems for the supertonic triad, D–F–A. While the minor triads on E and A (the mediant and submediant) produce the proportion 10:12:15, the minor triad on D is out of tune. Its fifth, D–A, is a syntonic comma flat, as is its third, D–F, which is in fact a Pythagorean minor third.

Tempering (increasing or decreasing) the size of intervals offered a practical solution to the problems inherent in just intonation by distributing the syntonic comma among the major thirds or the perfect fifths of the scale, thereby compromising the purity of one interval to preserve the purity of another. This practice became known as meantone temperament. Various systems of meantone temperament were put forward in the sixteenth and seventeenth centuries for the tuning of keyboard instruments, the most common of which was quarter-comma meantone temperament. In this system the perfect fifth is lowered by a quarter of a syntonic comma so that the major thirds have the pure ratio 5:4.

A perpetual problem with meantone temperaments is that, while modulation to closely related keys sounds pleasing, modulation to more remote keys sounds out of tune. The system of equal temperament, in which the

syntonic comma is distributed evenly among all twelve semitones of the octave, gradually became adopted because it removed the limitations on keys for modulation. The discrepancies between just and equal-tempered intervals are small and easily accepted by most listeners. The ratio of an equal-tempered semitone is $\sqrt[12]{2}$, or 1.05946...; by comparison, a just semitone, with ratio 16:15, is 1.06666.... The ratio of an equal-tempered perfect fifth, seven semitones, is $\sqrt[12]{2^7}$ or $\sqrt[12]{128}$, which is 1.498307..., whereas a just perfect fifth, with ratio 3:2, is of course 1.5. In equal temperament, one starts from a reference such as the note A, which is usually taken to have frequency 440 Hz.⁴ All other notes have frequencies of the form $440(\sqrt[12]{2})^n$, where n is the number of semitones between the note in question and the reference note A. In equal temperament, enharmonic notes such as C[#] and D^b are acoustically identical—that is, they share the same frequency. Equal temperament was well-suited for the kind of music that was written from the eighteenth century onward, with its much greater range of modulations and chromatic harmonic vocabulary.

The unit of the *cent* was defined by A. J. Ellis as the ratio between two pitches separated by one hundredth of an equal-tempered semitone, and became the most commonly used unit for measuring and comparing intervals.⁵ The octave consists, therefore, of 1200 cents. If a and b are two frequencies, then the distance in cents between the corresponding pitches is given by the formula $n = 1200 \log_2(a/b)$. (As a check, notice that if $a = 2b$ then one does indeed get the answer $n = 1200$.)

Microtonal systems based on the equal division of the octave into more than twelve parts were proposed and realized by some composers in the twentieth century, but they have not become widely used in Western music. However, the idea of dividing the octave into equal parts has become fundamental. It means that the notes used are naturally modeled by integers. If one regards two notes an octave apart as “the same,” which makes good musical sense, then one is dividing all notes into twelve EQUIVALENCE CLASSES [I.2 §2.3]. The natural model for these is arithmetic modulo 12.

4. The frequency of a pitch is a measurement of the number of cycles per second (abbreviated as “cps”). More commonly, the number of cycles per second is identified in units called *hertz* (abbreviated as “Hz”), named after the physicist Heinrich Rudolf Hertz.

5. Ellis’s account of the *cent* appeared in his appendix to the eminent nineteenth-century physicist Hermann von Helmholtz’s *On the Sensations of Tone* (1870; English edn., 1875).

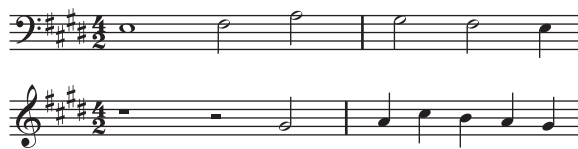


Figure 2 J. S. Bach, *Well-Tempered Clavier*, Book 2, Fugue no. 9, subject and diminution.

As we shall see later, the symmetries of the group of integers mod 12 are of great musical significance.

3 Mathematics and Music Composition

The association of number and music in acoustics was the result of scientific discovery. Number and music have also been associated through invention and creativity in music composition. Fundamental aspects of the temporal organization of music reflect simple proportional relationships. The basic durational values in Western music notation are the whole note (♩), half note (♪), quarter note (♫), eighth note (♬), etc. These are related to each other by simple multiples or fractions—all powers of 2—and these relationships are reflected in the metric organization of musical time into bars with the same number of beats. Bars or measures are indicated by time signatures such as the *simple meters* $\frac{2}{4}$, $\frac{3}{4}$, or $\frac{4}{4}$ (c), where beats (the ♩ in these examples) are typically subdivided into two, or the *compound meters* $\frac{6}{8}$, $\frac{9}{8}$, or $\frac{12}{8}$, in which beats (the ♪ in these examples) are subdivided into three.

A common device in musical composition, especially counterpoint, is for a melodic theme, or *subject*, to reappear at half or twice the original speed, techniques known as *rhythmic augmentation* or *diminution*, respectively. Figures 2 and 3 show the subjects of two fugues from the second volume of J. S. Bach’s *Well-Tempered Clavier*: no. 9 in E major, whose subject appears in diminution; and no. 2 in C minor, whose subject appears in augmentation. (The last note of the diminished or augmented subject may not be proportionally related to the original in order to allow a good continuation for the music that follows.)

Geometric relations have served as musical resources of other kinds too. A well-known construct in music theory is the *circle of fifths*, which was originally designed to demonstrate the relationships between different major and minor keys. As illustrated in figure 4, the twelve notes are arranged around the circle as a succession of perfect fifths. Any seven consecutive notes in this circle will be the notes of some major scale, which



Figure 3 J. S. Bach, *Well-Tempered Clavier*, Book 2, Fugue no. 2, subject and augmentation.

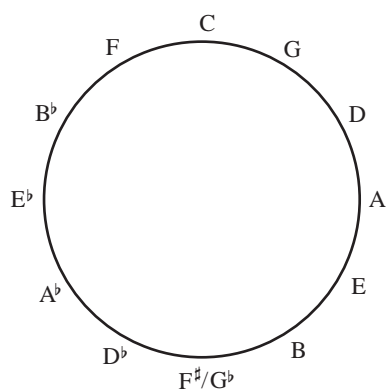


Figure 4 The circle of fifths.

makes it easy to understand some of the patterns of the key signatures. For instance, the C major scale consists of all the notes from F to B (clockwise). To change from C major to G major, one shifts the sequence by one, losing the note F but gaining F^\sharp . Continuing in this way, we see that C major is the key with no sharps or flats, G major has one sharp, D major has two sharps, A major has three sharps, etc. Similarly, moving counterclockwise from C, F major has one flat, B^\flat major has two flats, E^\flat major has three flats, etc. From a mathematical point of view, we have transformed the chromatic scale, which we identify with the additive group of integers mod 12, using the automorphism $x \mapsto 7x$, and this makes some musical phenomena much more transparent.

Reflective symmetry is another geometrical concept with a long history in musical composition. Musicians will frequently describe melodic lines in spatial terms, referring to notes of higher frequencies as “up,” and notes of lower frequencies as “down.” This allows one to think of melodic lines as ascending or descending. Reflection in a horizontal axis interchanges up and down. The musical counterpart to this is known as *melodic inversion*: one reverses the ascending or descending direction of each interval, and the result



Figure 5 J. S. Bach, *Well-Tempered Clavier*, Book 1, Fugue no. 23, subject and inversion.

is an inverted form of the melody. Figure 5 shows the subject of Fugue no. 23 in B major from the first volume of Bach's *Well-Tempered Clavier* and a later appearance of the subject in inverted form. A geometrical reflection is clearly visible in the notation, but, more importantly, the inversion can also be clearly heard in the sound of the music itself.

Conventional Western musical notation implies a two-dimensional organization, in which the vertical dimension expresses the relative frequency of pitches from low to high, and the horizontal dimension expresses chronological time from left to right. Another compositional device, much rarer than the devices of rhythmic augmentation and diminution or melodic inversion, is that of *retrograde*, where a melody is played backwards. When the melody is played backwards and forwards simultaneously, the technique is known as a *cancrizans* canon. Perhaps the best-known examples of cancrizans occur in the music of J. S. Bach, such as in the first canon of *The Musical Offering* or the first and second canons of the *Goldberg Variations*. Figure 6 shows the opening and closing measures of the cancrizans from Bach's *Musical Offering*. The melody of the first few bars of the upper staff returns in reverse order at the end of the piece in the lower staff, and likewise, the melody of the first few bars of the lower staff returns in reverse order at the end of the upper staff. Joseph Haydn's *Menuetto al rovescio*, from the Sonata no. 4 for violin and piano, is another well-known example of a similar technique, in which the first half of the piece is played backwards in the second half.

We may regard the devices of melodic retrograde and inversion as reflections in a two-dimensional musical space. However, retrograde is much more esoteric, owing to the greater constraints involved in the manipulation of musical time. Examples such as those by Bach and Haydn mentioned above demonstrate great ingenuity on the part of the composer, who must make the melodic retrogrades work convincingly with the underlying harmonic progressions. Certain common chord progressions, such as moving from the supertonic to



Figure 6 J. S. Bach, *The Musical Offering*, opening and closing measures of the cancrizans (canon 1).

the dominant, do not work well in reverse, so a composer attempting to write a cancrizans canon is forced to avoid them. Similarly, many common melodic patterns do not sound good when reversed. These difficulties account for the rarity of retrograde techniques in tonal music (i.e., music based on major and minor keys). With the abandonment of tonality in the early twentieth century, the main constraints were removed, making composition with retrograde easier. For instance, retrograde and inversion played an important role in serial music, as we shall see. However, composers of such music replaced the traditional constraints of tonal music with others, such as avoiding major or minor triads and bringing out other intervals deemed important for a particular piece.

The atonal revolution in the early twentieth century, during which composers experimented with novel methods of harmonic organization, led to the exploration of new types of symmetry relations in music composition. Scales based on repeating interval patterns (measured in semitones), such as the *whole-tone scale* (2-2-2-2-2-2) or the *octatonic scale* (1-2-1-2-1-2-1-2), appealed to composers for the symmetric structures and novel harmonies they embodied. The octatonic scale, also known in jazz circles as the *diminished scale*, had a particularly wide appeal among a variety of composers of different nationalities, such as Igor Stravinsky, Olivier Messiaen, and Béla Bartók. The novelty of the whole-tone and octatonic scales is that they have nontrivial *translational* symmetry, a property not shared by the major or minor scales. The whole-tone scale is unchanged if it is transposed by a tone, and the octatonic scale is unchanged if it is transposed by a minor third. There are thus only two distinct translates of the whole-tone scale and three of the octatonic scale. For this reason, neither scale has a clearly

defined tonal center, which was a major reason for their attractiveness to early twentieth-century composers.

Reflective symmetry was used by twentieth-century composers as well, to help them with the formal aspects of compositional design. A fascinating example is the first movement of Bartók's *Music for Strings, Percussion, and Celesta* (1936), which extends the traditional principles of the baroque fugue and incorporates a symmetric design. Figure 7 illustrates the structure of the fugue subject entries, starting from the initial entry on A. In a traditional fugue, the subject is stated in tonic, followed by a statement in the dominant, and then again in the tonic (and continuing the alternating pattern of tonic and dominant entries for fugues with more than three voices). In Bartók's fugue, the first statement of the subject begins with the note A, and the next with E. Instead of returning to A for the third statement, however, the subsequent entries follow a pattern of alternating fifths in opposite directions from A: that is, the sequence A-E-B-F[#], etc., alternates with the sequence A-D-G-C, etc. This pattern is illustrated in figure 7. Each of the interlocked cycles completes a circle of fifths, one clockwise, the other counterclockwise. Each letter in the illustration represents a statement of the fugue subject beginning on that note, and each of the interlocked cycles of fifths arrives on E^b (six semitones from the starting point, A) at its midpoint, so that all twelve notes occur once in the first half of the pattern and once again in the second half. The midpoint of the pattern corresponds to the dramatic climax of the work, after which the pattern of interlocked cycles of fifths resumes with subject entries in inverted form until the conclusion of the work with the return of the subject starting on A.

Arnold Schoenberg's twelve-tone method of composition, which he revealed in the early 1920s, is based on permutations of all the twelve notes, rather than of

Terri: Tim thinks this should stay as it is. OK?

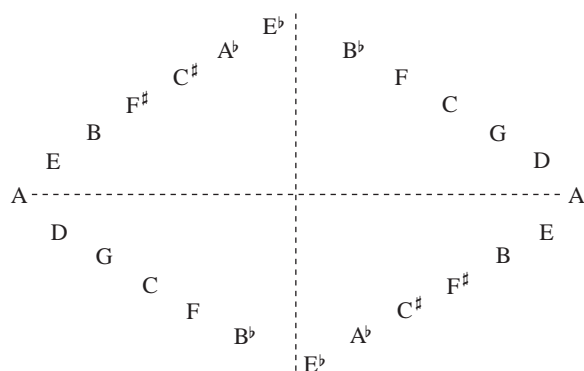


Figure 7 Plan of fugal entries in Béla Bartók's *Music for Strings, Percussion, and Celesta*, first movement (after Morris (1994, p. 61), with permission).

subsets of seven notes as one has in music in major or minor keys. In twelve-tone music (and atonal music in general), the twelve notes are supposed to have equal prominence: in particular, there is no single note with a special status like that of the tonic in a major or minor key. The basic ingredient of a piece of twelve-tone music is a *tone row*, which is a sequence given by some permutation of the twelve notes of the chromatic scale. (These notes can, however, be stated in any octave.) Once the tone row has been chosen, it can be manipulated by means of four types of transformation: transposition, inversion, retrograde, and retrograde inversion. Musical transposition corresponds to the mathematical operation of translation: the intervals between successive notes of a transposed row are the same as those between the corresponding notes of the original row, so the entire row is shifted up or down.⁶ Inversion corresponds to reflection, as we have already discussed: the intervals of the row are reflected about a "horizontal" axis. Retrograde corresponds to reflection in time: the row is stated backwards. (However, if it is combined with a transposition, as it may be, then it is better described as a glide reflection.) Retrograde inversion is a composition of two reflections, one vertical and one horizontal: it therefore corresponds to a half turn.

Figure 8 illustrates the serial transformations applied to a row created by Schoenberg for his *Suite for Piano*, opus 25, published in 1923. The forms of the row are

labeled P (for prime—the original row and its transpositions), R (for retrograde), I (for inversion), and RI (for retrograde inversion). The integers 4 and 10 in the row labels on the left and right refer to the starting notes of the P and I row forms by telling us how many semitones away from C they are. Thus, 4 refers to E (4 semitones above C) and 10 refers to B^b (10 semitones above C). The retrogrades of the P and I forms, R and RI, are labeled on the right-hand side of the figure. It is easy to see the inversion (reflection) of P₄ in I₄ about the first note E and the transposition of P₄ by six semitones in P₁₀, as well as the inversion of P₁₀ about the first note, B^b.

One may wonder what sort of insight we gain from understanding these abstract relationships and why they were so attractive to composers like Schoenberg. In Schoenberg's *Suite*, the eight row forms shown in figure 8 are in fact the only ones used in all five movements of the composition. This represents a high degree of selectivity, since there are 48 (= 12 × 4) available row forms. However, this self-imposed restriction is not on its own enough to account for the interest or attraction of this music. An additional aspect of the technique is that the row itself, and the way its transformations unfold in the course of a work, are chosen carefully to bring out certain relationships between notes. For example, all the row forms used in the *Suite* begin and end on the notes E and B^b, and these notes are frequently articulated in the work so that they take on an anchoring function that fills the void created by the absence of a conventional tonal center. Similarly, the notes in the third and fourth positions in each of the four row forms are always G and D^b, in either order, and likewise these are articulated in various ways in the movements of the *Suite* so that they can become recognizable. The two pairs of notes just mentioned, E–B^b and G–D^b, are related to each other by sharing the same interval, six semitones (half an octave, also known as the tritone because it spans three whole tones). In the hands of a master composer, a twelve-tone row is not a random collection of notes, but a foundation for an extended composition carefully constructed to produce interesting structural effects that one can learn to recognize and appreciate.

Permutations and serial transformations of other musical parameters besides pitch—such as rhythm, tempo, dynamics, and articulation—were explored by a new generation of postwar European composers, including Olivier Messiaen, Pierre Boulez, and Karlheinz Stockhausen. Compared with the serialization of pitch, however, serialization of these parameters does

6. Describing transposition as translation does justice to the fact that a melody sounds "the same" when transposed, even though the pitches are different, because the successive intervals are the same. If one arranges the twelve notes in a circle, then one can also think of this translation as a rotation.



Figure 8 Row forms in Schoenberg's *Suite for Piano* (1923).

not lend itself to such precise transformations, because it is less easy to organize them into discrete units than it is the twelve notes of musical space.

It is important to recognize that Schoenberg and most composers whose music exhibits mathematical conceptions such as those we have seen had little if any mathematical training.⁷ Nevertheless, the basic mathematical patterns and relations that we have discussed are so pervasive in so many aspects of so many different kinds of music that the importance of mathematics in music is undeniable.

We end this section with a few more examples. Proportional relations such as the simple ones between note values reappear on a larger scale in relations between lengths of formal divisions in music of Mozart, Haydn, and others: they often use basic building blocks of four-measure phrases and use them in pairs, and pairs of pairs, to form larger units. The techniques of melodic manipulation seen in Bach's works, which are found in a new guise in Schoenberg's twelve-tone techniques, can also be found in contrapuntal works of composers before Bach, such as Palestrina. And some composers, including Bach, Mozart, Beethoven, Debussy, Berg, and others, are said to have incorporated numerological elements into their composition,

such as symbolic numbers or proportions based on Fibonacci sequences and the golden ratio.

4 Mathematics and Music Theory

In the second half of the twentieth century, the ideas of Schoenberg and his followers were extended and developed in North American music theory. Milton Babbitt, a renowned American composer and theorist, is widely credited with introducing formal mathematics, specifically group theory, to the theoretical study of music. He generalized Schoenberg's twelve-tone system to any system where one has a finite set of basic musical elements (of which Schoenberg's twelve-tone rows were just one example), with relations and transformations between them (see Babbitt 1960, 1992). There are forty-eight ways of transforming a row, and Babbitt noted that these transformations form a group, which is in fact the product of the dihedral group D_{12} with the cyclic group C_2 of two elements. (The D_{12} in this product is the symmetry group of a dodecagon, and the C_2 allows the time reversal.) The four sets of transformations—P, I, R, and RI (see the previous section)—define a homomorphism from this group to the Klein group $C_2 \times C_2$, by identifying transformations that are equivalent up to rotation.

Identifying musical notes with the elements of the group \mathbb{Z}_{12} of integers mod 12, and modeling various musical operations by means of transformations on this group, makes it much easier to analyze some kinds of music, such as the atonal music of Schoenberg, Berg, and Webern, that do not lend themselves easily to more traditional analysis of harmony (see Forte 1973; Morris

7. Some composers, to be sure, have received extended mathematical training, which is reflected in their works. Iannis Xenakis, for example, was trained as an engineer, and had professional contact with the architect Le Corbusier. Xenakis found parallels between music and architecture through his study of Le Corbusier's *Modulor* system and its approach to form and proportion based on the human figure. Xenakis's compositions are characterized by their massive, physical sound and their complex algorithmic processes.

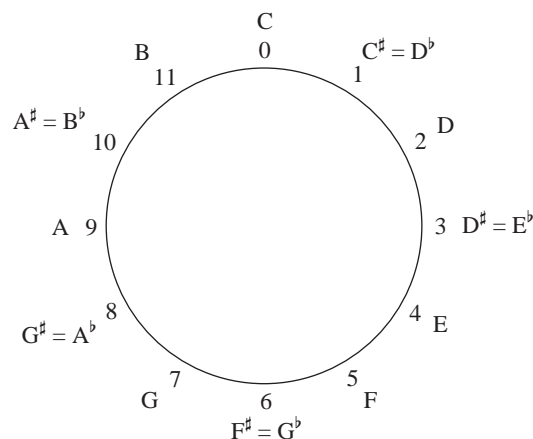


Figure 9 Circular model of the twelve notes (pitch classes).

1987; Straus 2005). This identification is illustrated in figure 9. As we have already commented, multiplying by 5 or 7 is an automorphism of \mathbb{Z}_{12} , and gives the cycle of fifths shown in figure 4 (when one substitutes the mod-12 integers for the names of the notes). This mathematical fact has many musical consequences. One of them is that it is common to substitute fifths by semitones, and vice versa, in chromatic harmony and in jazz.

A branch of music theory known as *atonal set theory* attempts to give a very general understanding of pitch relations by looking at all the $2^{12} = 4096$ possible combinations of notes, and defining two such combinations to be equivalent if one can be derived from the other by two simple transformations, the idea being that equivalent combinations will have the same intervals. The transformations in question are transposition and inversion. A transposition up by n semitones (where we think of n as an integer mod 12) is denoted T_n . The notation I is used for a reflection about the note C, so a general inversion takes the form $T_n I$ for some n . (Inversion in this context refers to reflection in musical space, and should not be confused with chord inversion in tonal music.) In these terms, to use a familiar example, the major triad and the minor triad are related to each other by inversion since their successive intervals are reflections of each other (four then three semitones in the major triad and three then four semitones in the minor triad, counting from the lowest note). Consequently, all major and minor triads belong to the same equivalence class. For example, the E-major triad $\{4, 8, 11\}$ is related to the C-major triad $\{0, 4, 7\}$ by the transposition T_4 (because $\{4, 8, 11\} \equiv \{0+4, 4+4, 7+4\}$, mod 12), and the G-minor triad $\{7, 10, 2\}$ is related by

inversion to the D-major triad $\{2, 6, 9\}$ by $T_4 I$ (because $\{7, 10, 2\} \equiv \{4 - 9, 4 - 6, 4 - 2\}$, mod 12). An equivalence class, such as the class of major and minor triads, will normally consist of twenty-four sets. However, if it has internal symmetries, such as those of the diminished seventh chord (with interval succession 3-3-3-3) or the whole-tone and octatonic scales mentioned earlier, then the number of sets in the class will be smaller, though it will always be a factor of 24.

Sets of notes in the same equivalence class share certain sonic attributes because they share the same number and types of intervals. But while it seems reasonable enough to regard transposed chords as equivalent, since they really do have an obvious “sameness” in the way they sound, there has been some controversy over the notion of inversional equivalence. For example, is it reasonable to declare major and minor triads to be equivalent to each other when they clearly do *not* sound the same and have very different musical roles? Of course, we are free to define any equivalence relation we like, so the real question is whether this one has any utility. And in some contexts it does: with sets of notes that do not possess extensive associations with tonal music it is easier to recognize this form of equivalence than it is with major and minor triads. For example, the three notes C, F, and B share the same intervals (one semitone, one perfect fourth or fifth, and one tritone) as the three notes F^\sharp , G, and C^\sharp , and this does indeed give them a noticeable form of “sameness.” (The set $\{11, 0, 5\}$ is inversionally related to $\{1, 6, 7\}$ by $T_6 I$ because $\{11, 0, 5\} \equiv \{6 - 7, 6 - 6, 6 - 1\}$, mod 12.)

There is other important work in music theory that has been inspired by group theory. The most influential example is David Lewin’s *Generalized Musical Intervals and Transformations* (1987), which develops a formal theory that connects mathematical reasoning and musical intuition. Lewin generalizes the concept of interval to mean any measurable distance, whether between pairs of pitches, durations, time points, or contextually defined events in a musical work. He develops a model called the generalized interval system (GIS), which consists of a set of musical objects (e.g., pitches, rhythmic durations, time spans, or time points), a group (in the mathematical sense) of intervals (representing the distance, span, or motion between a pair of objects in the system), and a function that maps all possible pairs of objects in the system into the group of intervals. He also uses GRAPH THEORY [III.34] to model musical processes, through his notion of a *transformation network*. The vertices of such a network are

basic musical elements such as melodic lines or chordal roots. These elements come with certain transformations, such as transposition (or shifting by a generalized interval) or the serial transformations from twelve-tone theory. Two vertices are joined by an edge if there is an allowable transformation that takes one to the other. The emphasis thus shifts from the basic elements to the relations that connect them. Transformation networks offer a dynamic way of looking at musical processes, giving visible form to abstract and often nonchronological connections in the analysis of musical works.

The level of generalization and abstraction makes Lewin's treatise a challenge for the mathematically unsophisticated music theorist, but it does not need more than fairly simple undergraduate-level algebra, so it is accessible enough for the determined reader with some mathematical training. It becomes clear to such a reader that the formality of the presentation is essential to a proper understanding of the transformational approach to music theory and analysis. Despite this formality, Lewin continually maintains contact with music itself, and how his mathematical tools can be applied in different contexts. The result is that the reader is rewarded with insights that would be impossible without the mathematical rigor. Mathematicians, while likely to find the material relatively elementary, may find their attention "captivated by the way in which the author gives new and, sometimes, unexpected interpretations to classical mathematical ideas when applied to musical contexts" (Vuza 1988, p. 285).

5 Conclusion

The playful Leibniz quotation with which this essay began underscores an enduring mathematical presence in music. Both disciplines rely in a fundamental way on concepts of order and reason, as well as more dynamic concepts of pattern and transformation. Music was once subsumed within mathematics, but it has now acquired its own identity as an art that has always derived inspiration from mathematics. Mathematical concepts have provided composers and theorists of music both with tools for creating music and with a language for articulating analytical insights about it.

Further Reading

- Babbitt, M. 1960. Twelve-tone invariants as compositional determinants. *Musical Quarterly* 46:246–59.
 ———. 1992. The function of set structure in the twelve-tone system. Ph.D. dissertation, Princeton University.

- Backus, J. 1977. *The Acoustical Foundations of Music*, 2nd edn. New York: W. W. Norton.
 Forte, A. 1973. *The Structure of Atonal Music*. New Haven, CT: Yale University Press.
 Hofstadter, D. R. 1979. *Gödel, Escher, Bach: An Eternal Golden Braid*. New York: Basic Books.
 Lewin, D. 1987. *Generalized Musical Intervals and Transformations*. New Haven, CT: Yale University Press.
 Morris, R. 1987. *Composition with Pitch-Classes: A Theory of Compositional Design*. New Haven, CT: Yale University Press.
 ———. 1994. Conflict and anomaly in Bartók and Webern. In *Musical Transformation and Musical Intuition: Essays in Honor of David Lewin*, edited by R. Atlas and M. Cherlin, pp. 59–79. Roxbury, MA: Ovenbird.
 Nolan, C. 2002. Music theory and mathematics. In *The Cambridge History of Western Music Theory*, edited by T. Christensen, pp. 272–304. Cambridge: Cambridge University Press.
 Rasch, R. 2002. Tuning and temperament. In *The Cambridge History of Western Music Theory*, edited by T. Christensen, pp. 193–222. Cambridge: Cambridge University Press.
 Rothstein, E. 1995. *Emblems of Mind: The Inner Life of Music and Mathematics*. New York: Times Books/Random House.
 Straus, J. N. 2005. *Introduction to Post-Tonal Theory*, 3rd edn. Upper Saddle River, NJ: Prentice Hall.
 Vuza, D. T. 1988. Some mathematical aspects of David Lewin's book *Generalized Musical Intervals and Transformations*. *Perspectives of New Music* 26(1):258–87.

VII.14 Mathematics and Art

Florence Fasanelli

1 Introduction

This article focuses on the relationship between the history of mathematics and the history of art in twentieth-century France, England, and the United States. The effect of mathematics on artists and the direct interactions between artists and mathematicians have both been extensively studied. These studies show that knowledge of mathematics has had a significant influence on many artists, as well as on musicians and writers. In particular, the increasingly wide acceptance, during the nineteenth century, of mathematical ideas that had once been revolutionary contributed strongly to what is now called modern art. At the end of the nineteenth century and the beginning of the twentieth, artists expressed on canvas and in sculpture their understanding of the fourth dimension and of NON-EUCLIDEAN GEOMETRY [II.2 §§6–10]. In doing so, they left behind their earlier training and heritage, which

had been heavily based on a mathematical perspective derived from EUCLID [VI.2]. Their new ideas reflected the progress that had been made in mathematics, and many of the artists who formed new schools of thought were also engaged in interpreting these new mathematical developments.

The connection between mathematics and art is rich, complex, and informative. This is evident in some of the artistic styles and the philosophies that developed under the influence of new mathematics (and science), and in the creation of mathematics to fulfill artistic needs. Some examples include the paintings (with their often-studied geometries) of Italian mathematician Piero della Francesca (ca. 1412–92), who, having made a transcription of Jacopo of Cremona's Latin translation of Archimedes' *Codex A*, wrote out his own mathematical theories of perspective; Hans Holbein's (1497–1543) *Ambassadors* (1533), which illustrates how an artist can use a distorted variation on mathematical perspective to fool the eye (anamorphosis); Artemesia Gentileschi's (1593–1652) deliberate correction of a smattering of blood in her first version of *Judith Beheading Holofernes* (1612–13) to a parabolic arc of blood in the second version (1620) to match a sketch that her friend, scientist, court mathematician, and amateur painter Galileo Galilei (1564–1642) had made as a study for his as yet unpublished law of projectile motion; various works of the Dutch portrait painter Johannes Vermeer (1632–75) using the camera obscura; Johann Hummel's (1769–1852) paintings of the making of the great granite bowl in Berlin, which used Gaspard Monge's (1746–1818) *Géométrie Descriptive* (1799); sculpture by Naum Gabo¹ (1890–1977) and his brother Antoine Pevsner (1886–1962) following their youthful academic study of solid geometry; and the mathematically understandable but physically implausible scenes by Maurits Cornelis Escher (1898–1972).

This article begins with a brief history of the development of perspective in art, because it is necessary to understand this in order to understand the rebellion against it that had such a decisive impact on modern art. This is followed by a short summary of the changing course of geometry in the nineteenth century through the development of non-Euclidean geometry and n -dimensional geometry. We then move on to the activities of artists, beginning in France in the early twentieth century and continuing with the works of

representative artists in other countries, all the while keeping in mind the mathematics that provoked their artistic responses.

2 Development of Perspective

During the fifteenth century artists were still primarily employed to produce images of sacred subjects, but there was an increased interest in having pictures match aspects of the physical world. Lacking any precursors, artists had to devise their own axioms of linear perspective. At the beginning of the sixteenth century these early ideas of mathematical perspective were spread by books that contained visual representations. Mathematics that was previously known only in writing or orally now took a visual form, which was copied in engravings and spread across Europe.

The first writings on perspective were by Leon Battista Alberti (1404–72) and Piero della Francesca, while the ideas of Filippo Brunelleschi (1377–1446), the Florentine architect and engineer who was in fact the first to consider a mathematical theory of perspective, were captured by his biographer Antonio Manetti (1423–97). Artists and mathematicians continued to develop the rules of perspective while looking for ways to best represent space and distance. Among the mathematicians, Federico Commandino (1509–75), renowned for his Latin editions of the works of Greek mathematicians such as Euclid, ARCHIMEDES [VI.3], and APOLLONIUS [VI.4], was the first to write about perspective for the benefit of mathematicians rather than artists. His student Guidobaldi del Monte (1505–1647) published the influential book *Perspectivae libri sex* in 1600, in which he showed that any set of parallel lines not parallel to the plane of the picture will converge to a vanishing point.

Great artists, notably Leonardo da Vinci (1452–1519) and Albrecht Dürer (1471–1528), were now portraying mathematics in a visual form. Mathematician Luca Pacioli's (1445–1517) *De Divina Proportione* (1509) includes Leonardo's unsurpassed woodcuts of polyhedra (among them the first published illustration of a rhombicuboctahedron), and Dürer's *Unterweysung der Messung* (1525) contains the first illustration of nets for models of polyhedra. Dürer's own new knowledge of perspective, whose secrets he had learned on a trip to Italy from Germany, inspired him to create his famous illustrations of how to draw a picture in which all the elements are in one-point perspective (see figure 1).

In the seventeenth century, Girard Desargues (1591–1661), a French engineer and architect who wrote on

1. Gabo was born Naum Neemia Pevsner but changed his name to distinguish himself from his painter brother.



Figure 1 An artist using Dürer's perspective machine.

practical subjects, continued the study of perspective that had been begun by the Renaissance artists. In doing so he invented a new, “non-Greek” way of doing geometry, which he published in his *Brouillon Project d'une Atteinte aux Événemens des Rencontres du Cône avec un Plan* (1639). In this essay, he attempted to unify the theory of conic sections through the use of projective techniques. This new PROJECTIVE GEOMETRY [I.3 §6.7] was based on his earlier realization that an artist can construct a perspective image without using a point from outside the picture field. However, of the original fifty printed copies of the *Brouillon Project* only one survives and his work, including his “perspective theorem,” was made known through the publications of other mathematicians. Abraham Bosse (1602–76), a friend of Desargues who ran a famous atelier where the art of engraving was taught, was responsible for publishing much of Desargues's work, including that on the theory of perspective. But Bosse's promotion of Desargues's innovative ideas created controversy in art circles and seriously damaged his professional reputation. However, in the twentieth century, when engraving was revived as an important art form, a replica of Bosse's studio was built in Paris.

In the early eighteenth century, the mathematician and amateur painter BROOK TAYLOR [VI.16] published *Linear Perspective: Or a New Method of Representing Justly All Manner of Objects as They Appear to the Eye in All Situations* (1715), the first book on perspective to give a general treatment of vanishing points. As Taylor wrote on the title page, the book was “a work necessary for painters, architects, etc., to judge of, and regulate designs by.” Taylor invented the phrase “linear perspective” and stressed the importance of what is now described as the main theorem of perspective: given any direction not parallel to the plane of the picture there is a “vanishing point” through which the representations of all lines in that direction must pass.

Since ancient times the axioms of Euclid's *Elements* have provided the basis for the understanding of two- and three-dimensional figures, and in the fifteenth century they provided the foundations for the study of perspective. But during the nineteenth century the long-standing debate about whether to accept Euclid's fifth axiom (the “parallel postulate”) was resolved in a way that was to provoke a radical change in the perception of geometry: it was demonstrated by several mathematicians—notably LOBACHEVSKII [VI.31] in 1829, BOLYAI [VI.34] in 1832, and RIEMANN [VI.49] in 1854—that a consistent “non-Euclidean geometry” was possible in which the fifth axiom no longer held.

The mathematician and expositor HENRI POINCARÉ [VI.61] provided popular accounts of these new ideas in his books *La Science et l'Hypothèse* (1902) and *Dernières Pensées* (1913), which were widely read in France and elsewhere. Poincaré's works provoked the highly influential French (and later American) artist Marcel Duchamp (1887–1968) to attach new meanings to the concepts of space and measurement. Duchamp famously discussed and used Poincaré's essays “Mathematical magnitude and experiment” and “Why space has three dimensions” to create artistic works of a completely new kind. (Duchamp's ideas have been explored by the art historian Linda Dalrymple Henderson, who has used Duchamp's extensive notes to analyze his understanding of four-dimensional and non-Euclidean geometries.)

3 Four-Dimensional Geometry

The modern movement known as *cubism* was greatly influenced by ideas of the fourth dimension. One of the ways cubists came into contact with these ideas, as well as with non-Euclidean geometry, was through their reading of popular science fiction. In *Gestes et Opinions du Docteur Faustroll* (1911), French author Alfred Jarry (1873–1907), a close friend of Spanish artist Pablo Picasso (1881–1973), attracted by the novelty of higher-dimensional geometries, wrote about the work of the British mathematician ARTHUR CAYLEY [VI.46]. In 1843, Cayley published “Chapters in the analytic geometry of n dimensions” in the *Cambridge Mathematical Journal*. This work, along with Hermann Grassmann's (1809–77) *Die Lineale Ausdehnungslehre*, published in German a year earlier, was of interest not only to mathematicians but also to the general public, who recognized that in spaces of higher than three dimensions basic concepts had to be redefined and generalized.

Terri: June says, “There is a connection as Jarry did draw on science fiction. I would be very against removing the whole reference to science fiction.” OK?

In 1880 Washington Irving Stringham (1847–1909), in another influential article, “Regular figures in n -dimensional space,” published in the *American Journal of Mathematics*, extended EULER’S FORMULA [I.4 §2.2] for polyhedra to new objects called “polyhedroids” in which polyhedra are joined by their faces so as to enclose a hyperspace. This article, which included illustrations of four-dimensional figures created by Stringham, was cited for the next twenty years in the most important mathematical texts on four-dimensional geometry. Stringham’s figures intrigued several artists during the first decade of the twentieth century: Albert Gleizes’s (1881–1953) painting *Woman with Phlox* (1910) has flowers that are similar to Stringham’s “*ikosatetrahedroid*”; while in Henri Victor Gabriel Le Fauconnier’s (1881–1946) *Abundance* (1910–11) Stringham’s “*hekatonikosihedroid*” appears.

Art forms evolved as artists found new ways of responding visually to the world around them. This was particularly true of cubism, in which the artist depicted objects from several viewpoints at once. In order to make sense of a cubist painting, the viewer was invited to construct a single (elusive) object from an array of different perspective “facets” laid out across the picture’s surface.

The n -dimensional geometries influenced not just the visual arts but also literature, including works by Rudyard Kipling and H. G. Wells, and music, for example Edgard Varese’s “*Hyperprism*” (1923). Some mathematicians used this new mathematics for humorous purposes: two examples were Charles Dodgson in his *Through the Looking Glass* of 1872 and Edwin Abbott in *Flatland: A Romance of Many Dimensions* of 1884. The latter in particular was read by French artists and was referred to in other mathematics books that they read, such as those by Esprit Pascal Jouffret (1837–1907).

4 Formal Protests against Euclid

In the early twentieth century, informed by Poincaré’s exposition of “the fourth dimension” and by knowledge of non-Euclidean geometry, a group of artists, including Gleizes and Jean Metzinger (1883–1956), explicitly attempted to liberate themselves from the geometry of three-dimensional Euclidean space. In an essay titled “*Du Cubisme*” they stated, “If we wished to tie the painter’s space to a particular geometry, we should have to refer to the non-Euclidean scholars; we should have to study, at some length, certain of Riemann’s theorems.” Here they appear to be referring to RIEMANNIAN GEOMETRY [I.3 §6.10], in which the notion of

shape is less rigid than it is in Euclidean geometry. They go on to say, “An object does not have one absolute shape, it has several, it has as many as it has planes in a range of meaning.” It is likely that they are referring here to Poincaré’s “*Les géométries non euclidiennes*” in *La Science et l’Hypothèse*. The title of a (lost) 1913 painting of Metzinger’s, *Nature morte (4^{me} dimension)*, gives a good indication of his interest in representing three and four dimensions on a two-dimensional surface. Both Riemann’s geometry and the fourth dimension lay behind what these artists were trying to accomplish; they referred to both, however, as “non-Euclidean.”

In 1918, enraged by the destruction wrought by World War I, a dozen artists, including Jean (Hans) Arp (1886–1986) and Francis Picabia (1879–1953), signed the *Dada Manifesto*. In it, they explicitly stated their belief that “all objects, sentiments, obscurities, apparitions and the precise clash of parallel lines are weapons for the fight [against conformity].” By the 1930s, more and more artists were using their knowledge of mathematics to change, in a radical way, the appearance of sculpture and painting.

5 Paris at the Center

During the last decade of the nineteenth century and the years before the outbreak of World War I, artists were profoundly influenced not just by mathematics, but also by the extraordinary developments and discoveries in science and technology. For instance, motion pictures (1880s), radios (1890s), airplanes, cars, X-rays (1895), and the discovery of electrons (1897) all had an impact on the work of artists. The pioneering painter Wassily Kandinsky (1866–1944) wrote that an artist’s block he was experiencing disappeared when he learned of what was new in science; his old world collapsed and he could begin painting again.

While it is not entirely clear how knowledge of scientific and mathematical thinking came to working artists in the early twentieth century, it is nevertheless evident that many artists were familiar with articles about mathematics written for the general public. There was also at least one tutor with whom they explored mathematics in depth. In 1911, in Paris, the mathematician and actuary Maurice Princet (1875–1971) gave informal lectures on four-dimensional geometry, using mathematician Esprit Pascal Jouffret’s *Traité Élémentaire de Géométrie à Quatre Dimensions et Introduction à la Géométrie à n Dimensions* (1903). Jouffret’s *Traité*, which makes reference to *Flatland*, contains ways to

Terri: spellings checked in Stringham’s original paper now, but because they’re very odd spellings we’ve stuck inverted commas in. All OK?

present four dimensions on paper, the diagrams by Stringham of polyhedroids in four-dimensional space, and clear presentations of the ideas and theories of Poincaré. A second book, *Mélanges de Géométrie à Quatre Dimensions* (1906), emphasizes similar points.

Princet's audience was the Puteaux cubist group (which was sometimes called the "Section d'Or"). The central figures in this group were the three brothers Raymond Duchamp-Villon (1876–1918), Duchamp, and Jacques Villon (born Gaston Emile Duchamp) (1875–1963). Princet's involvement with the artists continued, even after his divorce from Alice G  ry (1884–1975), who shared a bohemian life with the best man at their wedding, Pablo Picasso, and who later married Andr   Derain (1880–1954). G  ry had introduced Princet to the artists. An avid reader, she may have been the sitter for *Seated Woman with a Book* (1910), an early cubist painting by Picasso.

Together, in Paris, Princet and Duchamp privately studied Poincar   and Riemann, two important sources for Duchamp's work, as we have already seen. Duchamp's own notes, written a decade later as he created his famous painting *The Bride Stripped Bare by her Bachelors, Even (The Large Glass)* (1915–23), document his increasing interest in and understanding of four-dimensional and non-Euclidean geometries. Referring to Jouffret's book, which explained how a three-dimensional projection of a four-dimensional figure can be considered as a sort of "shadow," Duchamp told friends that the bride in his picture was a three-dimensional projection of a four-dimensional object recorded in two-dimensional form. He also refers to the fact, which fascinated him, that electrons were known to exist but could not be directly observed, claiming that his picture contained elements that were not directly represented. These notes, and others containing speculations on mathematics, were published in *   l'Infinif* (1966). Working in a field hitherto dominated by fifteenth-century Renaissance perspective and its dependence on a Euclidean framework, Duchamp and other artists learned with excitement that many mathematicians no longer felt it necessary to subject themselves to Euclidean restrictions, and art was dramatically changed.

Rather surprisingly, Riemann and Poincar   were even part of the original inspiration for Duchamp's famous "readymades," found objects presented as art. As the artist Rhonda Shearer described in the New York Academy of Science newsletter in 1997, Duchamp was very taken with Poincar  's description of the creative



Figure 2 Gabo's *Head No. 2*, COR-TEN steel, 1916 (enlarged version 1964). The works of Naum Gabo:    Nina Williams.

process in *Science and Method*. Poincar   reported on his accidental discovery of the so-called Fuchsian functions. After days of "unfruitful" conscious work spent trying to prove that the functions do not exist, he changed his habit and one evening drank black coffee late at night. The next morning, "fruitful" ideas came into his conscious mind. From these he selected "*tout fait*" (readymade) ideas and saw, surprisingly, a way to prove the existence of the mathematical functions whose existence he had previously doubted. Duchamp used the term "readymade" (and "*tout fait*") in 1915. The examples he selected, titled, and signed are ordinary manufactured objects such as a urinal turned upside down, *Fountain* (1917), and a bottle drying rack, *Bottle Rack* (1914), thought to be the first readymade.

6 Constructivism

In Russia in 1920, the artists Naum Gabo and Antoine Pevsner wrote that they had turned to mathematics in order to rethink their work. As they put it: "We construct our work as the universe constructs its men; as the engineer constructs his bridges; as the mathematician his formulas of the orbits." Gabo began to use a stereometric system that he had studied in engineering, creating sculptures such as "Head No. 2" (see figure 2). The subject of stereometry goes back at least as far as 1579, where it is listed in the "Groundplat" of John Dee's celebrated *Mathematicall Praeface* to Billingsley's edition of Euclid. It concerns the measurement of properties of solids, and was widely taught at universities



Figure 3 Gabo's two cubes: carving and construction.

in the nineteenth and twentieth centuries: indeed, it is still taught today in some European countries. Gabo and Pevsner constructed their sculptures out of planar parts, so space, rather than mass, became the sculptural element. Density was no longer important, with the result that the subtraction techniques used in classical sculpture (where material is carved away from a solid block leaving the artist's work as the solid) were no longer necessary. Sculpture became airy; surfaces became less significant and have remained so, at least within the tradition that became known as *constructivism*.

This tradition was first formalized in the Russian *Realistic Manifesto* (1920), written and signed by Gabo and Pevsner. There they argued that "The material formation of the object is to be substituted for its aesthetic combination. The object is to be treated as a whole ... a product of an industrial order like a car." Gabo took constructivism to the Bauhaus in Germany and then to France and England in the 1930s, where he worked alongside the British artists Barbara Hepworth (1903–75) and her husband Ben Nicholson (1894–1982). Gabo and Nicholson (with Leslie Martin) edited *Circle: International Survey of Constructive Art* (1937), which contained articles by themselves as well as ones by Hepworth, Piet Mondrian (1872–1944), and the critic Herbert Read (1893–1968), among others. In *Circle*, Gabo, referring back to the seventeen-year-old *Realistic Manifesto*, spells out what is meant by constructivism by guiding the reader to see how two cubes (shown in the photograph in figure 3) can illustrate the distinction between two kinds of representation of the same object: carving and construction. The cubes have different methods of execution and different centers of interest: one is mass and the other makes visible the space in which mass exists. Constructivism created an artistic context in which a mathematically understood space became a sculptural element. As Gabo wrote: "The stereometrical method in which [the right-hand cube] is executed shows elementarily the constructive principle of a sculptural space expression."

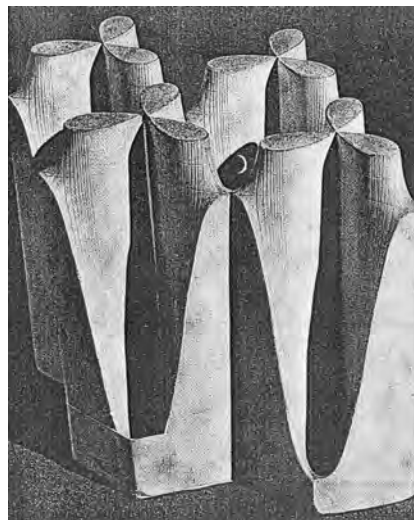


Figure 4 Man Ray's *Allure de la Fonction Elliptique*, 1936.

These artists studied mathematical models in museums and catalogues. These models, designed by mathematicians for teaching about surfaces, were made of string, cardboard, metal, and plaster. The same artists also studied photographs produced by the surrealist Man Ray showing strings and striations of surface lines on a model that had been found by another surrealist artist, Max Ernst, at the Institut Henri Poincaré in Paris. Ray portrayed these models with impressionistic patterns of light and shadow (see figure 4); he was interested in the "elegance"—the aesthetic persuasiveness—of the model, though aware that the original model-maker had sought to give visual form to an elegance inherent in the mathematical equations themselves. Other artists too, such as Hepworth and Gabo, stated that it was not mathematics itself but the beauty of the mathematical models that provided the inspiration for their work. Hepworth studied mathematical models that were on display in Oxford, considering them to be "sculptural working out of mathematical equations." They inspired her to add strings to her own work. However, she wrote that her inspiration was not the mathematics exhibited by the strings, but rather their power: "the tension I felt between myself and the sea, the winds, and the hills."

A close friend of both Gabo and Hepworth was the renowned sculptor Henry Moore (1898–1986). Moore too spoke and wrote about the influence of mathematical models on his work. He had seen stringed figures of Theodore Olivier (see figure 5) and after making many



Figure 5 Olivier's *Intersection of Two Hyperbolic Paraboloids*, 1830. Image courtesy of the Union College Permanent Collection, Schenectady, NY.



Figure 6 Moore's *Stringed Figure No. 1*, cherry wood and string, on oak base, 1937. Image (taken by Lee Stalsworth) courtesy of the Hirshhorn Museum and Sculpture Garden, Smithsonian Institution, Joseph H. Hirshhorn Purchase Fund (1989).

of his own mathematical models introduced strings into his sculpture in 1938, later considering it to have been the most abstract of his work. He said he "had gone to the Science Museum in South Kensington and had been greatly intrigued by some of the mathematical models ... hyperboloids and groins ... developed by [Fabre de] Lagrange in Paris, that have geometric fig-



Figure 7 Bill's *Eindeloze Kronkel*, bronze, 1953–56. Image courtesy of Mary Ann Sullivan, Bluffton University.

ures at the ends with colored threads from one to the other to show what the form between would be. I saw the sculptural possibilities of them, and I did some." Moore recognized that the use of strings connecting protrusions actually created a barrier between the solid sculpture and the space around the sculpture (see figure 6). The string barrier made it possible to see the captured space. Moore and Gabo made different uses of the mathematical models. As Moore later put it, Gabo "developed this string idea so that his structure always became space itself, whereas I liked the contrast between the solid and the strings ... I was making an outside shape a sculpture in its own right (Interior/Exterior forms), yet one which was not completed until each part was connected to the other."

7 Other Countries, Other Times, Other Artists

7.1 Switzerland and Max Bill

In the mid 1930s the Swiss designer and artist Max Bill (1908–94) became intrigued by a one-sided surface, unaware that it had been published in 1865 by the German mathematician and astronomer August Ferdinand MÖBIUS [VI.30]. Bill, when he needed a design for a sculpture to hang in a stairwell, independently invented his own MÖBIUS STRIP [IV.7 §2.3], by dangling a long narrow rectangle of flexible material and then attaching the corners appropriately (1935).

Having been informed some years later of the connection between his sculpture and its mathematical forerunner, Bill, who liked the simplicity of geometric forms, continued to earn commissions by making sculptures based on topological problems and single-sided surfaces (see figure 7). In a 1955 essay on the

mathematical approach in contemporary art, he wrote that mathematics, by giving all phenomena a meaningful arrangement, is an essential method to understand the world. For Bill, when mathematical relationships are given form they “emanate undeniable aesthetic appeal, such as goes out from space-models, as, for instance, those that stand in the Musée Poincaré in Paris.”

7.2 Holland and Escher

From the second half of the twentieth century onward there has been a groundswell of interest in the relationship between mathematics and art, particularly since 1992 when artists and mathematicians from around the world began holding joint annual conferences to explore old and new ideas about the connections between their disciplines. The popularity in the West of this interdisciplinary study is in no small part due to the unusual drawings and prints made by Maurits Cornelis Escher (1898–1972), a Dutch graphic artist—or “craftsman,” as he wished to be known. Escher was deeply interested in tessellations and “impossible” objects that are not constructible in three dimensions but that can nevertheless be portrayed in two dimensions. While his oeuvre is not thought of as an integral part of twentieth century art, he is greatly appreciated by mathematicians and also by the general public. Among his best-known works are pictures based on Penrose triangles and on the Möbius strip.

He was inspired by knowing and learning from mathematicians including Georg Pólya (1887–1985), Roger Penrose (1931–), and Harold Scott MacDonald “Donald” Coxeter (1907–2003). Escher was introduced to the international mathematics community in 1954 when the organizing committee for the Amsterdam meeting of the International Congress of Mathematicians inaugurated an exhibition of his work at the Stedelijk Museum. After Penrose viewed Escher’s 1953 print *Relativity* at this exhibition, he and his father, geneticist Lionel Penrose (1898–1972), were inspired to create impossible figures: the Penrose tribar and the Penrose staircase published in the *British Journal of Psychology* in 1958—the Penroses sent Escher an offprint of the article. Escher subsequently used these in two well-known lithographs: *Waterfall* (1961), in which water runs in perpetual motion from the base of a waterfall to the top of the waterfall; and *Ascending and Descending* (1960), which features a building with an impossible staircase which constantly rises or falls (depending on the direction you go around it) but returns to

the same level. Coxeter’s field was symmetry in the Euclidean and hyperbolic planes, but he also took pleasure in analyzing the works of artists from a mathematical point of view. Escher began a correspondence with him shortly after the congress, at which they met, and it lasted until his death in 1972. In 1957 Coxeter requested the use of two of Escher’s drawings to illustrate planar symmetry in “Crystal Symmetry and Its Generalizations,” his presidential address to the Royal Society of Canada—in this way Escher’s work spread among the mathematical community. In 1958, Coxeter sent Escher a letter containing a reprint of his address. The response was a request: “Could [you] give me a simple explanation how to construct the following circles, whose centers approach gradually from the outside till they reach the limit?” Coxeter’s reply, meant to be helpful, gave Escher one small piece of useful information; the rest of the lengthy letter was unintelligible to the artist. But from the pictures and his own keen geometric intuition, Escher was able to construct the circles he required, and by 1958 he was the first graphic artist to have used the three main geometries in his works: Euclidean, spherical, and hyperbolic. Coxeter was astounded that an artist, untrained in mathematics, could produce such accurate “equidistant curves” as he did in his 1958 woodcut *Circle Limit III*. Escher always claimed that he knew little mathematics, but many of his prints are a direct result of using mathematics. Mathematician Doris Schattschneider has said that Escher was really a “secret mathematician,” since much of his work depended on his pursuit of mathematical questions that arose from his interests and his interaction with mathematicians, which he referred to as “Coxetering.” He did, however, write that he preferred to find solutions and understanding by himself.

As well as his artistic and mathematical legacy, Escher had an important influence on crystallographers, who have used his symmetry drawings for analysis. Crystallographer Caroline MacGillavry has pointed out that Escher began a deep study of color symmetry and created a classification system in 1941–42, which was some time before crystallographers became interested in this field of study, which has become very active. The International Union of Crystallography subsequently commissioned Escher to illustrate MacGillavry’s *Symmetry Aspects of M. C. Escher’s Periodic Drawings*, first published in 1965. Its purpose was to interest “students in the laws which underlie repeating designs and their colorings.”

7.3 Spain and Dalí

As we have seen, some artists were influenced by their own knowledge of mathematics, others by a less direct appreciation of mathematical thinking, and still others by the appeal of mathematical models. Another kind of connection is illustrated by the example of the surrealist artist Salvador Dalí (1904–89) and his relationship with the mathematician and graphic artist Thomas Banchoff (1938–). Banchoff is a professor of mathematics at Brown University, known for his research in differential geometry in three and four dimensions. Since the late 1960s, he has also been involved in the development of computer graphics. Dalí's 1954 painting of Christ crucified on a hypercube was reproduced in a 1975 article about Banchoff's pioneering work, which used computer animation to illustrate geometry beyond the third dimension. This led to a series of meetings between Banchoff and Dalí over the next decade, at which hypercubes and other aspects of geometry and art were discussed. One joint project was the design for a giant sculpture of a horse that would appear realistic from only one viewing position. Dalí eventually envisioned a horse with its head in front of the viewer and its rump somewhere on the moon—clearly a project solely of the imagination. Dalí created works using anamorphoses, as other artists, beginning with Leonardo, had done. He prized his interactions with scientists and mathematicians, later stating, “Scientists give me everything, even the immortality of the soul.” Dalí also met the French mathematician René Thom (1923–2002) to discuss catastrophe theory, which, in 1983, he sought to represent in what turned out to be his last series of paintings.

7.4 Other Recent Developments: The United States and Helaman Ferguson

So far we have seen how mathematics has influenced art. Occasionally, artists have actually created mathematics, for instance to produce sculpture by means of carefully chosen mathematical equations. The noted American sculptor/mathematician Helaman Ferguson (1943–) divides his time equally between mathematics and the interpretation of mathematics in his art. As a mathematician he designs algorithms for operating machinery and for scientific visualization. In 1979 he found a method for finding integer relationships between more than two real or complex numbers—this was later named one of the top ten algorithms of the twentieth century. As an artist, he carves in stone. In



Figure 8 Ferguson's *Invisible Handshake II*: a triply punctured torus with negative Gaussian curvature.

1994, he asked mathematician Alfred Gray (1939–98) to develop equations for a Costa surface (named after the graduate student who invented equations for describing a minimal surface with holes), so that he could sculpt the surface (see figure 8). Gray developed the equations in terms of the Weierstrass zeta function. This could be used with Mathematica, which made it possible for Ferguson to create a stone sculpture. Ferguson sees his art as deriving from applied mathematics that has been developed over the course of the last two centuries:

Start with physical observations about soap films in nature (Plateau), write down a differential equation model describing minimizing surfaces (Euler-Lagrange), define a minimal surface geometrically in terms of curvature (Gauss), discover a minimal surface with non-trivial topology (Costa), draw computer images of the surface (Hoffman-Hoffman), recognize symmetry and prove the surface has not self intersections (Hoffman-Meeks), discover fast parametric equations for the surface (Gray), and finally return to nature with a sculpture, a solid form of a “soap film” big enough to touch and climb on.

7.5 The United States and Tony Robbin

The development of n -dimensional geometry also had a powerful effect on many other European and American artists, and this continued into the late twenti-

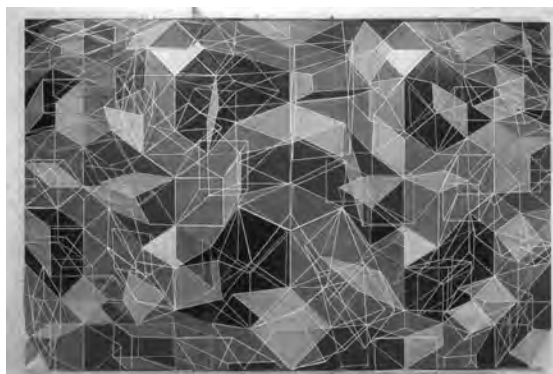


Figure 9 Robbin's *Lobofour*, acrylic on canvas with metal rods, 1982, collection of the artist.

eth century. Interest was boosted in the 1970s with the development of computer graphics by mathematicians and artists. Examples can be found in the work of American artist Tony Robbin (1943–), who has explored concepts of dimension in painting, prints, and sculpture (see figure 9). In late 1979, Robbin, who had also been a student of mathematics, was working on Banchoff's parallel processor computer and managed to visualize for the first time a four-dimensional cube, an event which radically changed his art, and which led him to develop two-dimensional works that portrayed the spatial fourth dimension. Writing in his book *Fourfield: Computers, Art & the 4th Dimension* (1992), Robbin tells us, "When the fourth dimension becomes part of our intuition our understanding will soar." Some of Robbin's constructions, paintings, and prints show figures in independent planes: that is, in overlapping spaces that cannot be fully seen in three dimensions. If the viewer wants to see two structures in the same place at the same time and rotating with respect to one another (as though projected from four-dimensional space), then looking at one of Robbin's wall-relief sculptures lit by red and blue light while wearing 3D glasses (one red and one blue lens) will create a full stereoscopic effect of the four-dimensional figure. In digital prints it is Robbin's lines and polyhedra that imply four dimensions, with the two-dimensional picture being a shadow of the higher-order object.

7.6 Hayter and Atelier 17

In 1927, the British surrealist and printmaker Stanley William Hayter (1901–88) decided to revive the almost lost skill of intaglio printing and established an experimental studio, "Atelier 17," in Paris. This was followed

by another in New York from 1940 to 1950 before he returned to Paris. Hayter was aware that many of the artists who used his facilities were working with a "different space from that seen through the classical window of Renaissance representation" that had existed when engraving flourished a hundred years earlier. The founding of Atelier 17 was central to the revival of the print as an autonomous art form, and Hayter's sensitivity to the significance of mathematics in the experimental techniques of printmaking (which had been evolving since the nineteenth century) is quite apparent: "Man's increasing consciousness of and power over space (in physics and mathematics) have been reflected in new and unorthodox methods of demonstrating space and time graphically," so that "many properties of matter and space, which had been represented diagrammatically only by the scientists, found their expression in graphic and affective forms." A printmaker in the twentieth century could use an arrangement of transparent webs to define planes above the picture plane. Specifically, by hollowing out spaces in the plate being engraved—possibly even gouging all the way to the bottom of the plate—the artist could make a projection in front of the plane of the picture. Although artists could have used this technique much earlier, it became important only at the end of the nineteenth century when the representational aspect of intaglio had been challenged by photography. They therefore used the gouge to create the third dimension. Hayter also describes in *About Prints* (1962) how Abraham Bosse's seventeenth-century atelier was organized and reconstructed in Paris in the twentieth century.

In World War II Hayter's interest in mathematics revealed itself in a more practical way, when, in collaboration with artist and patron of art Roland Penrose and others, he set up a camouflage unit and, as *Art News* reported in 1941, constructed

an apparatus which can duplicate the angle of the sun and the consequent length of cast shadows at any time of day, and day of the year, at any given latitude. This complex of turntables, discs inscribed with a scale of weeks, allowances for seasonal declination, and so on is just the kind of working mathematics he really delights in.

8 Conclusion

There has been a complex and fruitful relationship between Western art and mathematics in the twentieth century. Gabo, Moore, Bill, Dalí, and Duchamp are

notable artists who have been influenced by mathematics, and Poincaré, Banchoff, Penrose, and Coxeter are among the mathematicians who influenced them. In the other direction, twentieth-century mathematicians, like their forebears in the fifteenth and sixteenth centuries, often turned to art to explore and exhibit, or even just to explain more expressively, the meaning of their mathematics. They have also likened their creative processes to those of artists. As the French mathematician ANDRÉ WEIL [VI.93] wrote to his sister, author Simone Weil (1909–43), from military prison in 1940, “When I invented (I say invented, and not discovered) uniform spaces, I did not have the impression of working with resistant material, but rather the impression that a professional sculptor must have when he plays by making a snowman.”

Further Reading

- Andersen, K. 2007. *The Geometry of an Art: The History of the Mathematical Theory of Perspective from Alberti to Monge*. New York: Springer.
- Field, J. V. 2005. *Piero della Francesca: A Mathematician's Art*. Oxford: Oxford University Press.
- Gould, S. J., and R. R. Shearer. 1999. Boats and deckchairs. *Natural History Magazine* 10:32–44.
- Hammer, M., and C. Lodder. 2000. *Constructing Modernity: The Art and Career of Naum Gabo*. New Haven, CT: Yale University Press.
- Henderson, L. 1983. *The Fourth Dimension and Non-Euclidean Geometry in Modern Art*. Princeton, NJ: Princeton University Press.
- . 1998. *Duchamp in Context: Science and Technology in the Large Glass and Related Works*. Princeton, NJ: Princeton University Press.
- Jouffret, E. 1903. *Traité Élémentaire de Géométrie à Quatre Dimensions et Introduction à la Géométrie à n Dimensions*. Paris: Gauthier-Villars. (Digital reproduction available at www.mathematik.uni-bielefeld.de/~rehmann/DML/dml_links_title_T.html.)
- Robbin, T. 2006. *Shadows of Reality: The Fourth Dimension in Relativity, Cubism, and Modern Thought*. New Haven, CT: Yale University Press.
- Schattschneider, D. 2006. Coxeter and the artists: two-way inspiration. In *The Coxeter Legacy*, edited by C. Davis and E. Ellers, pp. 255–80. Providence, RI: American Mathematical Society.

Anne: I can confirm that the heading here is not any bigger than the previous part headings. OK?

Part VIII

Final Perspectives

VIII.1 The Art of Problem Solving

A. Gardiner

Where there are problems, there is life.

Zinoviev (1980)

In English the word “problem” has negative connotations, suggesting some unwanted and unresolved tension. Zinoviev’s reminder is therefore important: problems are the stuff of life—and of mathematics. Good problems focus the mind: they challenge and frustrate; they cultivate ambition and humility; they show up the limitations of what we know, and highlight potential sources of more powerful ideas. By contrast, the word “solving” suggests a *release* of tension. The juxtaposition of these two words in the expression “problem solving” may encourage the naive to think that this unwelcome tension can be massaged away by means of some “magic formula” or process. It cannot; there is no magic formula.

Why don’t we tell the truth? No one has the faintest idea how the process ... works, and in calling it a “process” we may be already making a dangerous assumption.

Gian-Carlo Rota, in Kac et al. (1986)

A “problem” is something that one wants to understand, to explain, or to solve, but which eludes one’s initial attempts to classify it as being of some familiar “type.” The experience of being confronted by such a “problem” is inevitably unsettling: it may eventually prove to be more familiar than one thought, but the would-be solver is initially dumped in terrain with few signposts or marked tracks. Some (such as Pólya and his recent followers) have tried to devise a universal “problem-solving meta-map.” But in reality there is no easy alternative to that painful immersion so familiar to generations of postgraduate students.

Grand general principles can help to make sense of this experience, but are unlikely to take us very far. Consider, for example, the four general principles formulated by DESCARTES [VI.11] in his *Discourse on Method*.

The first was never to accept anything for true which I did not clearly know to be such. The second, to divide each of the difficulties under examination into as many parts as possible, and as might be necessary for its adequate solution. The third, to conduct my thoughts in such order that by commencing with objects the simplest and easiest to know, I might ascend ... step by step to the knowledge of the more complex. And the last ... to make enumerations so complete ... that I might be assured that nothing was omitted.

Descartes’s rules are worth pondering. But it is hard to accept that it was the systematic application of these four rules that led to Descartes’s almost single-handed creation of analytic geometry as we know them today! In the detailed working out of the creative process, problem-specific “know-how” distilled from endless hands-on experience is likely to be far more important than any general principles. What then can one usefully say? To describe the “art of problem solving” in impressive-sounding detail would be irresponsible. But to say nothing would be misleading. Both options are unsatisfactory—yet these two responses are what students, teachers, and would-be mathematicians are most likely to meet! Attempts to teach “problem-solving” in schools often misconstrue mathematics as a kind of “subjective pattern-spotting.” Instead of correcting this distortion at university level, mathematicians often maintain a discreet public silence about the very private matter of *how* serious mathematical problems actually get solved. Hence, in addressing the theme for readers with a mathematical bent, this article has to start largely from scratch, and to proceed slowly. So we begin with a warning. The subject of problem solving is well worth exploring, but we shall proceed obliquely and

our conclusions will often remain implicit. Along the way we shall meet extracts from a number of sources—which may be viewed as an initial reading list for those who wish to pursue the theme in greater detail, provided they never forget that the only way to gain true insight into a craft is *through practicing the craft itself*. Mathematics may be “the queen of the sciences,” but the art of *doing* mathematics remains a *craft*, passed on in the ancient craft tradition, through painful initiation. A number of collections of problems at various levels—often using relatively elementary material—are listed in the references. Here we make do with a single example.

Problem. For all positive integers n and k , show that some triangular number is congruent to $k \pmod{2^n}$.

The reader is encouraged to explore this problem before reading on, noting any obvious stages along the way: from initial bewilderment, through an exploratory/organization phase, eventually culminating in a solution and an attempt to locate this isolated challenge in some broader mathematical context.

Mathematics is a largely unexplored “mental universe,” whose initial exploration and charting, subsequent colonization, routine traverse, and efficient administration correspond, in many ways, to the real-world adventures of geographical explorers in former centuries. To strike out beyond the security of the old-world coastline, to imagine and explore something new, takes intellectual courage.

Most prominent among these mathematical explorers are the “system builders,” who identify new mathematical continents, or who uncover profound and unexpected bridges joining known lands. Their initial motivation may stem from a specific problem, whose analysis provides hints of the outline of previously undiscerned structures; but the system builder’s focus then switches to the bigger picture: trying to identify, and to clarify, connections between the structures that underlie “mathematics in the large.” Such ventures often end up with little to show for them—they may come close to discovering some mathematical El Dorado, but they lack the gold to prove it. Some of these explorers may later be singled out as major prophets or discoverers, but such recognition can be fickle: those so honored may not have been the first to see their particular promised land; they may not have appreciated the significance of what they had stumbled upon, or of how it would eventually be seen to link known mathematical lands; their success may have depended on ear-

lier attempts by others; and their bounty may not have impressed their contemporaries as deeply as we now imagine.

Each triumph of the system builders is rooted in detailed knowledge of “mathematics in the small,” which may derive from work in a very different mathematical style—such as that of the mathematical beachcomber, who is most at home exploring the *known* mathematical shoreline, using some sixth sense to spot suspicious-looking rocks, under which are hidden intricate, and totally unexpected, microworlds on our very doorstep. While great explorers range further and further afield, they leave behind annoying gaps, or unsolved problems, which represent significant lacunas in our understanding—gaps that some future beachcomber may one day explain, so opening the way for some new synthesis.

The system builder and the beachcomber represent very different mental styles; but their contributions complement each other. In our evolving picture of the mathematical universe, insights on a small scale and on a large scale must somehow fit together. Hence the beachcomber’s chance discoveries may contribute in unexpected ways to our future conception of the large-scale mathematical universe.

Such differing styles should be borne in mind as we strive to make our introductory comments more specific. Our first attempt is based on a version of Alain Connes’s three levels of mathematical activity.

The first [level] is defined by the faculty of calculation—being able to apply a given algorithm rapidly and reliably.... The second level begins when the actual method of calculation is adapted to, and criticized in the context of, a particular problem.... In mathematics this is what often makes it possible to solve problems that aren’t too difficult or that don’t require any new ideas.... The third level [is] the level at which the mind, or rather conscious thought, is occupied with another task while the problem in question is being solved... subconsciously.... At [the third] level it isn’t only a matter of solving a given problem; it is also possible to discover... a part of mathematics to which the [previously] existing corpus gives no direct access.

Alain Connes, in Changeux and Connes (1995)

Connes’s first level focuses on the development of *robust technique*—that is, fluency, accuracy, and confidence in using given procedures in relatively standard ways. We say no more about work on this level except to stress its importance! Discussion about the “art of

problem solving” presupposes, and only makes sense in the context of, appropriate robust technique.

Connes’s second level includes most, but by no means all, of the serious mathematics that mathematicians engage in on a daily basis. Genuine problems occur on this level in different guises, ranging from (i) challenges designed to stretch the young would-be mathematician (in high school geometry, in puzzle books and problem-solving journals, in Olympiads, etc., whose material is designed to force the would-be solver to select, to adapt, and to combine known methods in unexpected ways), to (ii) genuine research problems that can be tackled and largely solved by selecting, adapting, and combining known methods in a suitably imaginative way.

In our problem about triangular numbers, the first level includes the immediate translation from words into symbols to obtain the congruence $m(m-1)/2 \equiv k \pmod{2^n}$, or $m(m-1) \equiv 2k \pmod{2^{n+1}}$, which, for arbitrary given $n \geq 1$, has to be solved for all $k \geq 1$. The second level might then include a systematic attempt to make sense of what happens for small values of n , leading to the formulation of simple conjectures whose proofs would solve the problem, followed by moves to devise the necessary proofs.

It is tempting to think of Connes’s third level as “in-scrutable,” in the spirit of the following extract:

In science, as well as in other fields of human endeavor, there are two kinds of geniuses: the “ordinary” and the “magicians.” An ordinary genius is a fellow that you or I would be just as good as if we were only many times better [than we are]. There is no mystery as to how his mind works. Once we understand what he has done, we feel certain that we, too, could have done it. It is different with the magicians. They are ... in the orthogonal complement of where we are and the working of their minds is for all intents and purposes incomprehensible. Even after we understand what they have done, the process by which they have done it is completely in the dark. They seldom if ever have students because they cannot be emulated and it must be terribly frustrating to cope with the mysterious ways in which a magician’s mind works.

Kac (1985)

However, one would then expect activity on this level to be so idiosyncratic as to be irrelevant to ordinary mortals. In fact, the most valuable insights we have into “the art of problem solving” derive from personal testimony about work on this level by precisely such “magicians” as POINCARÉ [VI.61], which suggests that there are clear parallels between the experience of the very best math-

ematicians on Connes’s third level and what happens when ordinary students, or mathematicians, operate “out of their depth” when tackling more mundane problems; that is, when their own fumbling requires them to work in regions to which *their own* “existing corpus gives no direct access.” In our problem about triangular numbers this might occur when a solver who has never met “congruences for binomial coefficients” manages to adapt the naive proof for $\binom{m}{2} \pmod{2^n}$ to cover the slightly more awkward $\binom{m}{3} \pmod{2^n}$, and realizes that, even though this naive approach does not extend to $\binom{m}{4} \pmod{2^n}$, something more general may be lurking in the darkness.

Thus we use the word “problem” to refer to *a serious mathematical challenge on at least Connes’s second level*, where this is to be interpreted in the spirit of activity on Connes’s second and third levels. So any analysis of the art of mathematical problem solving must somehow reflect experience on these two higher levels. By contrast, the educational assumptions that underpin most attempts to bring “problem-solving” to the classroom generally try to reduce this subtle process to a set of rules *in the spirit of Connes’s first level!*

A problem is much more than just a hard exercise. Consider the question, When is a “problem” *not* a problem? One answer is clearly, When it is too *easy*! However, many students and teachers are tempted to reject unfamiliar or mildly confusing problems because they appear to be *too hard*. This is an understandable reaction only where mathematics is limited to a succession of predictable exercises.

Most of us learn mathematics as a collection of standard techniques, which we use to solve standard problems in predictable contexts (Connes’s first level). Like the athlete or musician, the mathematics student needs to develop technique. However, as the athlete trains in order to *compete*, and the musician practices in order to *make music*, so the mathematician needs technique in order to *make mathematics* by tackling challenging problems. Each new piece of printed music may initially strike the beginner as a confusing array of black blobs. But as they work on the piece, phrase by phrase, it slowly takes on a shape of its own, revealing internal connections that may previously have been overlooked. Much the same is true when we confront an unfamiliar mathematical problem. At first sight we may not even understand the question. But as we struggle to make sense of the problem, we regularly find that, little by little, the fog begins to lift.

Terri: again, Tim thinks that this wording is the best available. OK?

Two rats fell into a can of milk. After swimming for a time one of them realized his hopeless fate and drowned. The other persisted, and at last the milk was turned to butter and he could get out.

In the first part of the war, Miss Cartwright and I got drawn into van der Pol's equation.... [W]e went on and on...with no earthly prospect of "results": suddenly the entire vista of the dramatic fine structure of solutions stared us in the face.

Littlewood (1986)

Terri: the publication that the quote(s) come from was published long after the work was done so the dates are correct. OK?

In 1923 HARDY [VI.73] and LITTLEWOOD [VI.79] made a conjecture about the number of arithmetic progressions (APs) of length k among the primes. One potential corollary was that the prime numbers must contain *arbitrarily long* APs. Faced with such a claim it is natural to start looking for APs which consist entirely of primes! But if you try, you will soon approach the limits of what is known: the first three odd primes, 3, 5, 7, form a very familiar AP of length *three*, but longer APs are surprisingly elusive (in 2004 the record for an AP of distinct primes had length twenty-three, with both the primes themselves and the step size being astronomical). Despite this unpromising lack of evidence, in 2004 Ben Green and Terence Tao proved that the set of prime numbers does indeed contain arbitrarily long APs. Their proof is a fine example of the way in which significant progress often combines a detailed reevaluation of known results (in this case a deep result of Szemerédi), lateral thinking (they embed the primes not in the integers, but in a natural but sparser set of "almost primes" of which the primes constitute a *nonzero* fraction), and the determination and ingenuity to make such ideas deliver the goods.

It remains a serious challenge to capture the essence of Littlewood's experience (where the fog suddenly lifts) in a form that is suitable for relative beginners, whether through time-constrained problems (see Barbeau 1989; Gardiner 1997; Lovasz 1979), or through structured investigations (see Gardiner 1987; Ringel 1974). In the year in which Green and Tao announced their proof, the British Mathematical Olympiad posed the following problem, which readers are encouraged to tackle.

Problem. In an AP of seven distinct primes, what is the smallest possible value of the largest prime?

This challenge could enliven any introductory number theory course, as well as providing a natural link to recent developments. For the novice it is far from obvious how to begin, but the basic idea is elementary and

should be "known" (in some sense), and can be used to generate natural APs of lengths 4, 5, 6, 7, 8, provided that one accepts the value of carrying out extensive computations quickly and intelligently.

A great discovery solves a great problem. But there is a grain of discovery in the solution of any problem. Your problem may be modest; but if it challenges your curiosity and brings into play your inventive faculties, and if you solve it by your own means, you may experience the tension and enjoy the triumph of discovery. Such experiences at a susceptible age may create a taste for mental work and leave their imprint on mind and character for a lifetime.

From the preface to the first printing of Pólya (2004)

Pólya is, if anything, too reticent here. The important distinction is not between that which is "known" and that which is truly "original," but rather between mathematical activity in the spirit of Connes's first level and mathematical activity in the spirit of Connes's second and third levels. Any introduction to this distinction is inevitably through problems whose solution is *known to someone*, so we should collect and use good "modest problems," not apologize for them. Ulam puts it more directly.

I learned chess from my father.... The moves of the knight fascinated me, especially the way two enemy pieces can be threatened simultaneously with one knight. Although it is a simple stratagem, I thought it was marvelous, and I have loved the game ever since.

Could the same process apply to the talent for mathematics? A child by chance has some satisfying experiences with numbers; then he experiments further and enlarges his memory by building up a store of experiences.

Ulam (1991)

Children also find delight—if less profound and more short-lived—in the discovery that one can set up a "corner move" in the children's game noughts and crosses (tic-tac-toe) so as to *simultaneously* threaten to complete two lines-of-three, at most one of which can be countered. This delight in a double-edged strategy, which points in two directions at once, has much in common with the pleasure we derive from (i) puns and *double entendres* in ordinary language, in humor, and in poetry, (ii) the almost physical response when we recognize subtle variations on a theme in music, and (iii) the more cerebral appreciation we feel when we meet counting methods based on unanticipated isomorphisms, or the essentially two-faced idea of "proof

by contradiction” in mathematics. This enjoyment of hidden ambiguities and double meanings is related to the evident (but poorly understood) way in which *analogy* guides, and delights, mathematicians of all ages.

Banach once told me, “Good mathematicians see analogies between theorems or theories; the very best ones see analogies between analogies.”

Ulam (1991)

Koestler, in his thought-provoking book *The Act of Creation* (1976), shows how scientific and literary “creativity” often flows from the identification and exploitation of “double meanings with a built-in tension.” (Koestler calls them *bisociations*: “the perceiving of a situation or idea *L* in two self-consistent but habitually incompatible frames of reference ... the event *L* is made to vibrate simultaneously on two different wavelengths, as it were.”) His study begins with an analysis in precisely this vein of the human response to humor, both comic and tragic, including a selection of jokes attributed to VON NEUMANN [VI.91]!

Ulam’s innocent-sounding question (in the extract before last) challenges us not only to provide children with “satisfying experiences with numbers,” but also to identify other quintessential aspects of mathematics and to ensure that they are experienced memorably at school (and undergraduate) level. In particular, insofar as there is such a thing as an “art of problem solving,” we need to learn how to convey it faithfully and effectively through the medium of classical elementary mathematics to those who are near the beginning of their mathematical studies, or who may not yet have any commitment to mathematics.

It is often claimed that Pólya’s little book *How to Solve It* provides an answer. It does not. Pólya was a pioneer who sought to provoke a debate among mathematicians about “heuristics.” This debate never really got started. Instead his first low-level attempt at a theoretical framework has been embraced uncritically.

Much of what Pólya writes about specific problems in *How to Solve It* makes sense; but his general conclusions on “how to help students solve problems” are less convincing. As a result, much of the book’s general theorizing needs to be read extremely carefully. For example, Pólya’s suggestion that “when the teacher solves a problem before the class, he should dramatize his ideas a little and he should put to himself the same questions which he uses when helping students” is spot on. But alarm bells should start ringing when he

confidently concludes that “[thanks] to such guidance, the student will eventually ... acquire something that is more important than the knowledge of any particular mathematical fact.” In the right setting the claim may occasionally be true; but as a statement about the effect on students in general it is false.

Similar claims have been widely used to justify the introduction of a whole new branch of school mathematics called “problem-solving” (see NCTM (1980) and www.pisa.oecd.org), which has grown *at the expense of* mastery of the “particular mathematical facts” on which the activity itself depends.

Pólya and others were right to insist that school mathematics should include a regular diet of good problems, and that educators have a duty to convey not just the techniques and inner logical structure of the subject, but also the experience of struggling to uncover the mathematics hidden in multistep problems and carefully structured investigations. Fortunately, the four volumes that Pólya wrote to illustrate this broader thesis remain in print (Pólya 1981, 1990). There the focus is on mathematics, and the rhetoric is more restrained:

[L]et us learn proving, but also *let us learn guessing*.... I do not believe there is a foolproof method to learn guessing. At any rate, if there is such a method, I do not know it, and quite certainly I do not pretend to offer it in the following pages.... [P]lausible reasoning is a practical skill and it is learned, as any other practical skill, by imitation and practice.

Pólya (1990, volume 1)

These four books should be compulsory reading for all serious mathematics educators, graduate students, and mathematics lecturers. However, Pólya and others failed to show how problem solving could be developed *within* the standard school mathematics curriculum. Instead they concentrated on proposing general rules that might “help students become better problem solvers.” What is needed is to clarify (i) which aspects of elementary mathematics have the potential to captivate young minds—not because they are more “enjoyable” in some superficial sense, but because they are more “pregnant with meaning”; and (ii) *how to teach* such material so as to convey this deeper meaning on an elementary level. This is not the place for a detailed analysis, but we suspect such an analysis would *strengthen* the position of many traditionally important topics and themes, encouraging them to be taught in such a way as to bring out their inherent richness, while recogniz-

ing that these goals depend on prior mastery of certain basic techniques without which this richness can scarcely be appreciated. In contrast, recent “reforms,” whose declared intention was to *enrich* school mathematics, have regularly *reduced* both the emphasis on, and the time available for, serious elementary mathematics.

Those who want good problems to enrich school mathematics often fail to recognize that well-intentioned “reforms” are usually unstable under the kind of distortions that routinely affect large-scale educational change (where the cultivation of professional competence, sensitivity, independence, and responsibility among teachers is regularly replaced by centralized control via a fragmented list of separate “outcomes,” which are then assessed in ways that actively discourage good teaching).

Small-scale experiments can also have unintended side effects! As a little-known example of a radical attempt to cultivate the art of problem solving at school level we offer Eisenstein’s account of his own education at lower secondary school (1833–37).

[E]ach student had to prove the theorems consecutively. No lecture took place at all. No one was allowed to tell his solutions to anybody else and each student received the next theorem to prove, independent of the other students, as soon as he had proved the preceding one correctly, and as long as he had understood the reasoning.... While my peers were still struggling with the eleventh or twelfth, I had already proved the hundredth.... [T]his method ... can probably not be adapted.... One does not obtain that overview of the whole subject, which can only be achieved by a good lecture.... In the end, the best mathematical genius cannot discover alone what has been discovered by the collaboration of many outstanding minds.... For students this method is only practicable if it deals with small fields of easily understandable knowledge, especially geometric theorems, which do not require new insights and ideas.

Eisenstein (1975)

Eisenstein was a remarkable mathematician. Yet at the tender age of twenty, on the threshold of the mathematical world that he longed to inhabit, he could see the limitations of this approach—even for students such as himself.

Problems that cultivate a taste for problem solving tend to incorporate certain characteristic features, such as simplicity, rhythm, naturalness, elegance, and surprise; and their solutions are often double-edged. But their most important feature is that, while their solu-

tion should be within reach of those in the target audience, the statement of the problem should convey no direct hint as to how to begin. Indeed, a good problem may continue to frustrate the would-be solver for a disturbingly long period.

A tacit rite of passage for the mathematician is the first sleepless night caused by an unsolved problem.

Reznick (1994)

The role of sleep and sleeplessness in creative problem solving is well documented (if poorly understood). It often features within the “incubation” phase of HADAMARD’s [??] “four phases” (discussed below), which summarize the process through which the initial experience of helplessness and leaden frustration is sometimes transmuted into golden success.

Such success is neither mechanical, nor the result of pure chance. In solving a good problem—as with a good puzzle—there is no magic problem-solving method that might relieve us of the need to struggle: the struggle may sometimes be fruitless, but it is an important part of the process. Thus, a successful outcome generally presupposes a certain kind of preparatory hard work. When asked how he made his discoveries, GAUSS [VI.26] is said to have answered, “Durch planmässiges Tattonieren,” that is, through systematic and persistent groping around!

Having discovered a way into a problem, one may realize that it “should have been obvious” where to begin; but things are often obvious only in retrospect. One learns by experience how a certain kind of persistence can cause the fog that initially surrounds an unfamiliar problem to magically evaporate; what was at first invisible then stands out so clearly that one can scarcely understand how it could ever have been missed.

When faced with an unfamiliar mathematical problem, the mathematician, young or old, is like someone who is trying to open some fiendishly difficult Chinese puzzle box with a hopelessly small bunch of keys. At first glance the surface seems totally smooth, without a single visible crack. If you were not convinced that it was indeed a Chinese puzzle box, and that it could in fact be opened, you would soon give up. Knowing (or rather *believing*) that it can be opened, you may be willing to keep searching until you eventually begin to discern the slightest hint of a crack here and there. You may still have no idea how the pieces are meant to move, or which of your “keys” may help you to open up the first layer of the puzzle, but by trying the most appropriate-looking keys in the most promising cracks,

you eventually stumble on one that fits exactly, and the pieces begin to move. The job is certainly not done; but the mood has changed and you feel you are well on the way.

As we have already seen, this experience of initial confusion, giving way as one grapples with a problem to unexpected insight, is in no way confined to beginners. It is part of the very nature of mathematics and of the way human beings do mathematics. If a problem is unfamiliar, its solution may require persistence, faith, and much time. So one should never give up too easily, and should always be prepared to look back after solving a problem to see what one could perhaps have done differently.

It is most important in creative science not to give up. If you are an optimist you will be willing to “try” more than if you are a pessimist. It is the same in games like chess. A really good chess player tends to believe (sometimes mistakenly) that he holds a better position than his opponent. This, of course, helps to keep the game moving and does not increase the fatigue that self-doubt engenders. Physical and mental stamina are of crucial importance in chess and also in creative scientific work.

Ulam (1991)

Persistence is of course easier to sustain if one has a degree of optimism about the likely outcome, or if one has cultivated the sheer “bloody-mindedness” that makes one refuse to give up (as with Littlewood’s surviving rat). However, there are dangers.

I learned, subconsciously, from Mazur how to control my inborn optimism and how to verify details. I learned to go more slowly over intermediate steps with a skeptical mind and not to let myself be carried away.

Ulam (1991)

At the International Congress of Mathematicians in Paris in 1900, HILBERT [VI.63] presented twenty-three major research problems, which he judged would be important for the development of mathematics in the twentieth century. These problems seemed very hard; yet in bringing them to the attention of his fellow mathematicians Hilbert felt the need to stress that this should not be used as an excuse for putting off trying to solve them.

However unapproachable these problems may seem to us and however helpless we stand before them, we have, nevertheless, the firm conviction that their solution must follow by a finite number of purely logical processes.... This conviction of the solvability of every

mathematical problem is a powerful incentive to the worker. We hear within us the perpetual call: There is the problem. Seek its solution. You can find it by pure reason; for in mathematics there is no *ignorabimus*.

During the nineteenth century it became clear that the more that scientists discovered about nature, the more they realized *how little they knew*, and that one could never hope to discover “the whole truth.” This realization was summed up by the physiologist Emil du Bois-Reymond in the phrase “*ignoramus et ignorabimus*”—ignorant we are and ignorant we shall remain. As the new century dawned, Hilbert felt that it was important to state as clearly as he could that *mathematics is different*. In mathematics, he said, we can tackle problems with “the firm conviction that their solution must follow by a finite number of purely logical processes.” As if to underline his assertion, one of his problems was solved almost immediately (though the most famous, the RIEMANN HYPOTHESIS [IV.2 §3], remains unresolved).

Hilbert was talking about mathematical *research*: but his principle applies even more strongly when tackling problems from textbooks, Olympiads, or university courses. When faced with an unfamiliar and apparently very difficult mathematical problem, we have little choice about how to proceed: we must either tackle the problem using the “bunch of keys” or mathematical techniques that we already know (no matter how limited they may be), or put off trying. Of course it is important to learn new tricks, and to revise old ones, as we go along. And of course there is always the temptation to imagine that the problem we face is simply *too* hard, that progress toward a solution requires some trick or technique that we have not yet learned and that the solution is therefore beyond our powers. This defeatist view is all the more plausible because it must sometimes be true! Mathematicians know perfectly well that, strictly speaking, the assumption that every problem can be solved is irrational (in that it cannot be justified logically, and is in general clearly false: we now know that some problems are intrinsically insoluble as stated). *It is nevertheless an invaluable working hypothesis*. Thus we should never let such doubts interfere with the basic hypothesis that *every problem we tackle has to be solved using essentially the techniques that we already know* (deployed with sufficient ingenuity!). Though strictly illogical, the assumption that every problem can be solved has justified itself so often in practice that it becomes a powerful conviction—a conviction that is psychologically invaluable each time we

experience that feeling of helplessness when trying to get to grips with a hard mathematical problem.

Hilbert's judgment that his problems would play a central role in the mathematics of the twentieth century was remarkably astute. But the most interesting thing for us here is his rallying call: however unapproachable these problems may seem at first sight, and however helpless we stand before them, we have the firm conviction that their solution must be possible by purely logical processes. "There is the problem. Seek its solution. You can find it by pure reason." As in most printed mathematics, Hilbert offered no psychological guidance on how to proceed. Those who took up Hilbert's challenge were expected to discover such things for themselves.

Like every social activity, mathematics has a "front" and a "back": the *front* is where the finished products are displayed for public consumption, while the *back* is where the real work is done in less presentable surroundings. A naïve realist might view the *front* as a mere facade, insist that all serious "problem solving" goes on "out back," and declare this separation to be artificial.

Sometime, in a future that is knocking at our door, we shall have to retrain ourselves and our children to properly tell the truth. The exercise will be particularly painful in mathematics. The enrapturing discoveries of our field systematically conceal, like footprints erased in the sand, the analogical train of thought that is the authentic life of mathematics.... Until that day, however, the truths of mathematics will make only fleeting appearances, like shameful confessions whispered to a priest, to a psychiatrist, or to a wife.

In the nineteenth chapter of "The Betrothed," Manzoni describes as follows the one genuine moment in a conversation between astute Milanese diplomats: "It was as if, between acts in the performance of an opera, the curtain were to be raised too soon, and the spectators were given a glimpse of a half-dressed soprano screaming at the tenor."

Gian-Carlo Rota, in Kac et al. (1986)

However, the prospect of some mathematical equivalent of being obliged to witness "a half-dressed soprano screaming at the tenor" should cause us to hesitate before embracing Rota's vision of the future.

The *front-back* metaphor is due to the sociologist Erving Goffman. One standard example is that of a restaurant. We tend to think of a restaurant in terms of what we see "out front," where the manners, food, and language are "all dressed up"; but everything we see out front is totally dependent on the raw heat, the

steam and grease, the conflicts and curses "out back" in the kitchen—where the hard work is done to tight deadlines and in very different conditions.

The triumph of mathematics in the modern world has been largely due to the fact that these two worlds—the front and the back—have been deliberately and systematically separated. It may seem curious that we have no agreed way of discussing the dynamics of the mathematical kitchen; but mathematics has grown largely because its practitioners have learned to separate its *objective* results, and the way they are validated and presented, from the intriguing, but inscrutable (and ultimately irrelevant!) *subjective* alchemy through which these mathematical results are conjured up. This formal separation has led to the adoption of a universally communicable format, which transcends personal taste and style, and which can therefore be comprehended, checked, and improved by anyone. Any move to pay greater attention to the mental, physical, and emotional dynamics that underlie mathematical problem solving must understand the need for this separation and respect the formal world of "objective" mathematics.

There are intriguing insights into the human dynamics of the mathematical kitchen scattered throughout the mathematical literature. One such insight is the fact that different mathematicians may have very different styles, even though most of these differences are rarely discussed. One example is the perceived role of *memory*. Some mathematicians value memory highly.

It seems to me that a good memory—at least for mathematicians and physicists—forms a large part of their talent. And what we call talent or perhaps genius itself depends to a large extent on the ability to use one's memory properly to find the analogies, past, present and future, which, as Banach said, are essential to the development of new ideas.

Ulam (1991)

Others have an excellent memory for anything *within their own field of interest*, but have considerable difficulty storing information from outside that domain in an easily retrievable form. And many would-be mathematicians are drawn to the subject precisely because they see it as requiring markedly less memorizing than most other disciplines. The important point would seem to be not *how much* one remembers, but *what* one makes automatic, and *how accessibly* this and other information is stored. It is clearly worth making a serious effort to organize in one's mind that material which

Terri: Tim says that this is a standard juxtaposition in analytic philosophy. OK?

is central to one's own work—so that it is available for instant use. It is also important, as we shall see, to collect a penumbra of possibly useful ideas, information, and examples—so that the mind is in a position to make incidental connections which might be fruitful. But it is not necessarily wise to learn in a uniform way everything that might conceivably be needed for the problem at hand: knowing slightly less sometimes forces the mind to get by on less, and hence to be more ingenious or inventive.

Hadamard's Four Phases

Littlewood's (1986) numerous perceptive observations concerning his contemporaries highlight other differences in style—such as speed and working habits. Similar insights may be found in many of the livelier mathematical autobiographies, but Littlewood's remarks are especially valuable.

With a good deal of diffidence I will try to give some practical advice about research and the strategy it calls for. In the first place research work is of a different "order" from the learning process of pre-research education (essential as it is). The latter can easily be rote-memory, with little associative power: on the other hand, after a month's immersion in research the mind knows its problem as much as the tongue knows the inside of one's mouth. You must acquire the art of "thinking vaguely", an elusive idea that I can't elaborate in short form.... I should stress the importance of giving the subconscious every chance. There should be relaxed periods during the working day, profitably I say spent in walking.

Littlewood (1986)

At one stage Poincaré thought there might be just two main styles of mathematical thinking:

The one sort are above all preoccupied with logic.... The other sort are guided by intuition and at the first stroke make quick but sometimes precarious conquests.... [O]ne often says of the first that they are *analysts* and calls the others *geometers*.

Poincaré (1904)

But in identifying the label "logical" with that of "analyst," and the label "intuitive" with that of "geometer," he noticed that HERMITE [VI.47] constituted a counterexample—an "intuitive analyst"! Clearly, the range of mathematical styles is more complex (see Hadamard 1945, chapter VII). One consequence is that any analysis of the art of problem solving in general needs to be drawn with a broad brush. Despite this

caveat, Hadamard's "four-phase" model of mathematical creativity has found widespread acceptance, so it may help if one's work habits respect these phases:

It is usual to distinguish four phases in creation: preparation, incubation, illumination and verification, or working out.... Preparation is largely conscious, and anyhow *directed* by the conscious. The essential problem has to be stripped of accidentals and brought clearly into view; all relevant knowledge surveyed; possible analogues pondered. It should be kept constantly before the mind during intervals of other work.... Incubation is the work of the subconscious during the waiting time, which may be several years. Illumination, which can happen in a fraction of a second, is the emergence of the creative idea into the conscious. This almost always occurs when the mind is in a state of relaxation and engaged lightly with ordinary matters.... Illumination implies some mysterious rapport between the subconscious and the conscious, otherwise emergence could not happen. What rings the bell at the right moment?

Littlewood (1986)

Pólya's *How to Solve It* proposes a less convincing four-stage "recipe" for the problem-solving process ("understand, plan, act, reflect"), which has nevertheless been widely used at school level. Hadamard's four phases provide a useful framework for thinking and communicating about the creative process; they also separate the relatively routine aspects (which one may be able to influence more easily) from the more elusive ones. The "conscious *preparation*" phase is perhaps the most mundane stage, requiring a combination of method and discipline. Littlewood again offers sound advice. He recognizes that his advice may not suit all tastes, but he insists that we would all benefit from trying different patterns of working in order to identify and cultivate habits that are as effective as possible.

Most people need half an hour or so before being able to concentrate fully.... The natural impulse towards the end of a day's work is to finish the immediate job: this is of course right if stopping would mean doing work all over again. But try to end in the middle of something; in a job of writing out, stop in the middle of a sentence. The usual recipe for warming up is to run over the latter part of the previous day's work; this dodge is a further improvement.... When I am working really hard I wake around 5.30 a.m. ready and eager to start; if I am slack I sleep till I am called.

Littlewood (1986)

At some ill-defined stage, this preparation achieves a sufficiently clear understanding of the immediate prob-

Terr: again the proofreader marked up a change to US punctuation here (adding a serial comma) but this is a quote from a UK publication. Keep as it is?

Terr: proofreader suggested transposing quote marks and comma here, but as the quote comes from a CUP publication, I expect (although don't know) that the punctuation appeared like this. I don't mind transposing but wanted your opinion before I did so.

lem, together with a level of saturation in relevant background information, to enable the mind to begin trying different approaches and combinations of ideas. We have reached the *incubation* phase.

We cannot know all the facts, since they are practically infinite in number.... Method is precisely the selection of facts.

Poincaré (1908)

I've often observed too that once the first hurdle of preparation has been surmounted, one runs up against a wall. The main error to be avoided is trying to attack the problem head-on. During the incubation phase you have to proceed indirectly, obliquely.... Thought needs to be liberated in such a way that subconscious work can take place.

Alain Connes, in Changeux and Connes (1995)

Temperament, general character, and "hormonal" factors must play a very important role in what is considered to be purely "mental" activity.... A "subconscious brewing" (or pondering) sometimes produces better results than forced, systematic thinking.... [W]hat we call originality... might to some extent consist of a methodical way of exploring all avenues—an almost automatic sorting of attempts....

When I remember a mathematical proof, it seems to me that I remember only salient points, markers, as it were, of pleasure or difficulty. What is easy is easily passed over because it can be reconstructed logically with ease. If, on the other hand, I want to do something new or original, then it is no longer a question of syllogism chains. When I was a boy I felt that the role of rhyme in poetry was to compel one to find the unobvious because of the necessity of finding a word which rhymes. This forces novel associations and almost guarantees deviations from routine chains or trains of thought. It becomes paradoxically a sort of automatic mechanism of originality.... What people think of as inspiration or illumination is really the result of much subconscious work and association through channels in the brain of which one is not aware at all.

Ulam (1991)

It takes two to invent anything. The one makes up combinations; the other one chooses, recognizes what he wishes and what is important to him in the mass of the things which the former has imparted to him. What we call genius is much less the work of the first one than the readiness of the second one to grasp the value of what has been laid before him and to choose it.

Paul Valéry, quoted in Hadamard (1945)

We have reached a double conclusion: that invention is choice [and] that this choice is imperatively governed by the sense of scientific beauty.

Hadamard (1945)

Part of the pleasure (and pain), the magic (and masochism) of mathematics stems from the fact that the next step—from incubation to illumination—remains so mysteriously elusive. *Illumination* can occur at any time. In most cases—especially where the realization is of something relatively straightforward—this occurs during periods of "official work." However, this need not be so, especially when the corner to be illuminated is especially dark or unfamiliar, or if the leap of imagination required is large. In such cases it seems that, after the hard graft of the preparation and incubation phases, the mind often needs to "step back" in order to see the way forward more clearly. That is, hard work needs to be combined with relaxation, as Connes implies when he warns against "trying to attack the problem head-on." In one oft-quoted example, Poincaré recalls how he realized the profound connection between Fuchsian functions and hyperbolic geometry as he stepped aboard a bus while on a day out! The first three extracts below show that the mind may achieve this in-between state as a result of *sleeplessness*, or *in the very act of waking*. The fourth extract concerns strenuous *hill walking*. What is common to them all is that the moment of enlightenment does not occur while the beneficiary is officially working!

It was his custom to tell his friends that if others would meditate as long and as deeply as he did on mathematical truths, they would be able to make his discoveries. He said that often he meditated for days on a piece of research without finding a solution, which finally became clear to him after a sleepless night.

Dunnington (1955)

One phenomenon is certain and I can vouch for its absolute certainty: the sudden and immediate appearance of a solution at the very moment of sudden awakening. On being very abruptly awakened by an external noise a solution long searched for appeared to me without the slightest instant of reflection on my part... and in a quite different direction from any of those which I had previously tried to follow.

Hadamard (1945)

Most striking at first is this appearance of sudden illumination, a manifest sign of long, conscious prior work.... The role of this unconscious work in mathematical invention appears to me incontestable....

For a fortnight I had been attempting to prove that there could not be any function analogous to what I have since called Fuchsian functions. I was at that time very ignorant. Every day I sat down at my table and spent an hour or two trying a great number of combinations, and I arrived at no result. One night I took some black coffee, contrary to my custom, and was unable to sleep. A host of ideas kept surging in my head; I could feel them jostling one another, until two of them coalesced, so to speak, to form a stable combination. When morning came, I had established the existence of one class of Fuchsian functions, those that are derived from the hypergeometric series. I had only to verify the results, which only took a few hours.

Poincaré (1908)

I had been struggling for two months to prove a result I was pretty sure was true. When ... walking up a Swiss mountain, fully occupied by the effort, a very odd device emerged—so odd that though it worked I could not grasp the resulting proof as a whole.... I had a sense that my subconscious was saying, “Are you *never* going to do it, confound you; try this.”

Littlewood (1986)

The resulting sense of satisfaction is familiar even to those whose mathematical experience is limited.

Illumination is not only marked by the pleasure—the exhilaration!—one inevitably experiences at the moment it strikes, but also by the relief one suddenly feels at seeing a fog abruptly lift, and disappear.

Alain Connes in Changeux and Connes (1995)

However, after months of hard work, such intoxication can sometimes be deceptive.

In mathematics one cannot stop at drawing with a big, wide brush; all the details have to be filled in at some time.

Ulam (1991)

The *verification*, or working-out, process often appears mundane; but it is rarely routine, and regularly reveals hidden subtleties that force us to reassess the anticipated approach. Unforeseen difficulties may remain unresolved, and we may be obliged reluctantly to begin the cycle all over again. It is tempting to think of this as “failure.” But mathematics is not a mere machine for solving problems; it is a way of life. In their different ways success and failure both send us back to the drawing board—as Gauss observed in a letter to BOLYAI [VI.34] in 1808.

It is not knowledge, but the act of learning, not possession but the act of getting there, which grants the greatest enjoyment. When I have clarified and exhausted a

subject, then I turn away from it, in order to go into darkness again; the never-satisfied man is so strange—if he has completed a structure, then it is not in order to dwell in it peacefully, but in order to begin another. I imagine the world conqueror must feel thus, who after one kingdom is scarcely conquered, stretches out his arms for others.

Further Reading

- Barbeau, E. 1989. *Polynomials*. New York: Springer.
- Changeux, J.-P., and A. Connes. 1995. *Conversations on Mind, Matter, and Mathematics*. Princeton, NJ: Princeton University Press.
- Dixon, J. D. 1973. *Problems in Group Theory*. New York: Dover.
- Dunnington, G. W. 1955. *Carl Friedrich Gauss: Titan of Science*. New York: Hafner. (Reprinted with additional material by J. J. Gray, 2004. Washington, DC: The Mathematical Association of America.)
- Eisenstein, G. F. 1975. *Mathematische Werke*. New York: Chelsea. (English translation available at <http://www-ub.massey.ac.nz/~wwiims/research/letters/volume6/>.)
- Engel, A. 1991. *Problem-Solving Strategies*. New York: Springer.
- Gardiner, A. 1987. *Discovering Mathematics: The Art of Investigation*. Oxford: Oxford University Press.
- . 1997. *The Mathematical Olympiad Handbook: An Introduction to Problem Solving*. Oxford: Oxford University Press.
- Hadamard, J. 1945. *The Psychology of Invention in the Mathematical Field*. Princeton, NJ: Princeton University Press. (Reprinted 1996.)
- Hilbert, D. 1902. Mathematical problems. *Bulletin of the American Mathematical Society* 8:437–79.
- Kac, M. 1985. *Enigmas of Chance: An Autobiography*. Berkeley, CA: University of California Press.
- Kac, M., G.-C. Rota, and J. T. Schwartz. 1986. *Discrete Thoughts: Essays on Mathematics, Science, and Philosophy*. Boston, MA: Birkhäuser.
- Koestler, A. 1976. *The Act of Creation*. London: Hutchinson.
- Littlewood, J. E. 1986. *A Mathematician's Miscellany*. Cambridge: Cambridge University Press.
- Lovasz, L. 1979. *Combinatorial Problems and Exercises*. Amsterdam: North-Holland.
- NCTM (National Council of Teachers of Mathematics). 1980. *Problem Solving in School Mathematics*. Reston, VA: NCTM.
- Newman, D. 1982. *A Problem Seminar*. New York: Springer.
- Poincaré, H. 1904. *La Valeur de la Science*. Paris: E. Flammarion. (Contained in *The Value of Science: Essential Writings of Henri Poincaré* (2001), translated by G. B. Halsted. New York: The Modern Library.)
- . 1908. *Science et Méthode*. Paris: E. Flammarion. (Contained in *The Value of Science: Essential Writings of Henri Poincaré* (2001), translated by F. Maitland. New York: The Modern Library.)

Terri: your advice would be appreciated about the format of this reference and the next. Different people translated different bits of Poincaré's original work for this volume. Also, is the year correctly set? I couldn't find a CMS instruction about what to do in a situation like this. Thanks.

- Pólya, G. 1981. *Mathematical Discovery*, two volumes combined. New York: John Wiley.
- . 1990. *Mathematics and Plausible Reasoning*, two volumes. Princeton, NJ: Princeton University Press.
- . 2004. *How to Solve It*. Princeton, NJ: Princeton University Press.
- Pólya, G., and G. Szego. 1972. *Problems and Theorems in Analysis*, two volumes. New York: Springer.
- Reznick, B. 1994. Some thoughts on writing for the Putnam. In *Mathematical Thinking and Problem Solving*, edited by A. H. Schoenfeld. Mahwah, NJ: Lawrence Erlbaum.
- Ringel, G. 1974. *Map Color Theorem*. New York: Springer.
- Roberts, J. 1977. *Elementary Number Theory: A Problem Oriented Approach*. Cambridge, MA: MIT Press.
- Ulam, S. 1991. *Adventures of a Mathematician*. Berkeley, CA: University of California Press.
- Yaglom, A. M., and I. M. Yaglom. 1987. *Challenging Mathematical Problems with Elementary Solutions*, two volumes. New York: Dover.
- Zeitz, P. 1999. *The Art and Craft of Problem Solving*. New York: John Wiley.
- Zinoviev, A. A. 1980. *The Radiant Future*. New York: Random House.

VIII.2 “Why Mathematics?” You Might Ask

Michael Harris

It seems to me that they have a poor opinion of our religion if they think it needs the protection of philosophy.

Lorenzo Valla, *Dialogue on Free Will*

1 A Metaphysical Burden

ANDRÉ WEIL [VI.93], speaking at the 1978 International Congress of Mathematicians at Helsinki, concluded his address entitled “History of Mathematics: Why and How?” with these words:

Thus my original question “Why mathematical history?” finally reduces itself to the question “Why mathematics?,” which fortunately I do not feel called upon to answer.

Proceedings of the ICM, Helsinki, 1978
(pp. 227–36, quotation on p. 236)

I heard Weil’s address, and the applause that followed, and remember imagining circumstances in which that final question could not be so easily evaded. For instance, in 1991 the House Committee on Science, Space, and Technology called upon the American Mathematical Society (AMS) to answer a very similar question: “What are the main goals in the mathematical sciences?” Weil knew his audience, and the committee of

twelve mathematicians responding to the government body responsible for research budgets knew theirs:

The most important long-term goals for the mathematical sciences are: provision of fundamental tools for science and technology, improvement of mathematics education, discovery of new mathematics, facilitation of technology transfer, and support of efficient computation.¹

“Meaning is what makes things sell,” wrote Roland Barthes (1967), and the AMS adopted the posture of FOURIER [VI.25], who, according to a celebrated comment of JACOBI [VI.35], included in a letter to LEGENDRE [VI.24] of July 2, 1830,

... had the opinion that the principal aim of mathematics was public utility and explanation of natural phenomena; but a philosopher like him should have known that the sole end of science is the honor of the human mind.

It might seem that the AMS has left a place for “honor” in its third goal, but a later elaboration of that goal directs the reader toward “unexpected” applications of pure mathematics.

Few pure mathematicians are as indifferent to practical applications as HARDY [VI.73], who in *A Mathematician’s Apology* famously claimed that: “Judged by all practical standards, the value of my mathematical life is nil.” But it is fair to assume that, when they are addressing one another rather than government committees, most pure mathematicians (including those who represented the AMS in 1991) would choose a quite different list of “most important long-term goals.”

In this they have long been able to count on the protection of philosophy. It has been a commonplace since Plato to grant mathematics intrinsic value on metaphysical grounds.² The topos of mathematics as a source of certain knowledge was already well established by the second century, when Ptolemy wrote

1. From “Pilot assessment of the mathematical sciences (prepared for the House Committee on Science, Space, and Technology),” *Notices of the American Mathematical Society* 39 (1992):101–10.

2. The present essay is mainly concerned with metaphysical certainty. Descartes wrote in *Principles of Philosophy* (chapter CCVI) of “certainty ... founded on the metaphysical ground that, as God is supremely good and the source of all truth, the faculty of distinguishing truth from error which he gave us, cannot be fallacious so long as we use it aright, and distinctly perceive anything by it,” and cites “the demonstrations of mathematics” as his first example. Plato (in *Republic*, VII, 522–31) saw mathematics rather as a source of “knowledge of that which exists forever.” Certainty and its cognates are some, but only some, of the apparent blessings of mathematics that so impressed certain philosophers as to “infect” the whole of their work, as Ian Hacking (2000) argues.

Only mathematics, if one attacks it critically, provides for those who practice it sure and unswerving knowledge, since the demonstration comes about through incontrovertible means, by arithmetic and geometry.³

THE CRISIS IN THE FOUNDATIONS OF MATHEMATICS [II.7] of the early twentieth century, which culminated in GÖDEL'S INCOMPLETENESS THEOREMS [V.18], was largely motivated by a desire to make mathematical certainty safe from dependence on human frailty. As RUSSELL [VI.71] wrote in *Reflections on my Eightieth Birthday*:

I wanted certainty in the kind of way in which people want religious faith. I thought that certainty is more likely to be found in mathematics than elsewhere.... Mathematics is, I believe, the chief source of the belief in eternal and exact truth.

Quoted in Hersh (1997)

Russell's hope to ground certainty in logic is largely a thing of the past—as Marvin Minsky wrote in another context, "without an intimate connection between our knowledge and our intentions, logic leads to madness, not intelligence" (Minsky 1985/1986)⁴—but his words continue to echo. After Jean-Pierre Serre was named first recipient of the Abel Prize, he was quoted in *Libération* (May 23, 2003) to the effect that mathematics is the only producer of "totally reliable and verifiable" truths. And Landon T. Clay III, announcing the creation of the \$7 000 000 Millennium Prize Fund, linked his decision to devote much of his personal fortune to the support of pure mathematics to "the decline in religious certitude... the pursuit of proof continues to be a strong motivating force in human actions."⁵

The mind saves its honor, as Jacobi would have it, but only through indenture to a higher power. I would like to express my opinion that the bargain implicit in comments like those just quoted, placing pure mathematicians on the front lines in defense of metaphysical certainty or some other normative concern of philosophers, is an unnecessary burden that fails to do justice to what is uniquely valuable about mathematics. It also

fails to protect pure mathematics from the real existential dangers it faces, of which budget cuts are only the most obvious expression. Mathematics is not likely to collapse for lack of a coherent account of its certainty, but it may well collapse for lack of an account of its value.

2 Postmodernism versus Mathematics?

One danger that should not worry mathematicians is that of *postmodernism*. Many thousands of pages have been written on this topic, although it is not clear that the word designates anything specific. I will nevertheless add a few pages of my own, because the term has come to be used as shorthand for a radical relativism that is thought to call into question not only certainty but rationality in all its forms.⁶ One thus finds mathematicians who are skeptical of certainty in Russell's sense but who nonetheless express hostility to something they call "postmodernism" as they try to defend reason and the value of mathematics as a rational activity.

Applied to architecture, postmodernism designates a reasonably precise tendency. As a trend defining the spirit of the times, it has been called "the cultural logic of late capitalism," differing from modernism by emphasizing space rather than time, multiple perspectives and fragmentation rather than unity of meaning and totality, pastiche (sampling)⁷ rather than progress, and much more along the same lines. As a movement in philosophy it is most typically (if abusively) associated with Michel Foucault, Jacques Derrida, Gilles Deleuze, Roland Barthes, Jean-François Lyotard, and more generally the "French theory" of the 1960s and 1970s. Postmodern prose is eclectic, ironic, self-referential, and hostile to linear narrative. The variant known as posthumanism celebrates the fading of conceptual and material boundaries between human beings and machines.

We are all postmodernists insofar as we have experienced the degradation of public discourse under the influence of advertising slogans, and are therefore likely, in spite of ourselves, to read Jacobi's invocation of "the honor of the human mind" as a precursor of that genre. Mathematicians can even claim to be

3. See Lloyd (2002), in which is cited Ptolemy's *Syntaxis*, I, chapter 1, 16.17–21.

4. Compare René Thom's comment in connection with his criticism of attempts to reduce mathematics to set theory: "In attempting to attach meaning to all the phrases constructed in ordinary languages, according to Boolean rules, the logician proceeds to a phantasmic, delirious reconstruction of the universe" (reprinted in Tymoczko 1998).

5. Transcript of interview by Francois Tisseyre conducted on the occasion of the Paris Millennium Meeting, May 24, 2000, graciously provided by the Clay Mathematics Institute.

6. For example, Lakoff and Núñez (2000) write of a "radical form of postmodernism which claims that mathematics is purely historically and culturally contingent and fundamentally subjective." No examples are given of texts espousing this point of view.

7. "Because his...artistry comes from combining other people's art,...the DJ is the epitome of a postmodern artist" (www.jahsonic.com/postmodernism.html).

the first postmodernists: compare an art critic's definition of postmodernism—"meaning is suspended in favor of a game involving free-floating signs"—with the definition of mathematics, attributed to HILBERT [VI.63], as "a game played according to certain simple rules with meaningless marks on paper."⁸ Mathematics could nevertheless (or for that very reason) safely ignore postmodernism, were it not that the latter is supposed to have no room for certainty, metaphysical or otherwise.⁹ So it is not surprising that authors who are considered postmodernists have had some perplexing run-ins with science and mathematics.

The typically controversial postmodernist account of science sounds like this:

Science and philosophy must jettison their grandiose metaphysical claims and view themselves more modestly as just another set of narratives.

Terry Eagleton's caricature of postmodernism,
quoted in Harvey (1989)

As far as mathematics is concerned, relativism of this kind has more to do with English-language postmodernism than with the French original. One might have thought that mathematical progress from axioms to theorems and from lesser to greater abstraction or generality constituted a prime example of the sort of "master narrative" that French postmodernists regarded with suspicion, and a particularly tempting target given the special role Enlightenment thinking reserved for mathematical explanation; but that seems not to have been the case. Although the most prominent French philosophers associated with postmodernism were metaphysical skeptics in other regards, they had no quarrel with mathematics' metaphysical pretensions per se; but they did question their relevance to the human sciences. For Derrida, thinking of LEIBNIZ [VI.15] in particular, "[mathematics] was always the exemplary model of scientificity" (in *Of Grammatology*, p. 27), and Foucault claimed that:

Mathematics has certainly served as a model for most scientific discourse[s] in their efforts to attain formal

rigor and demonstrativity; but for the historian who questions the actual development of the sciences, it is ... an example ... from which one cannot generalize.¹⁰

The Archeology of Knowledge (pp. 188–89)

At least one of postmodernism's canonical French texts does take on the issue of certainty in science and mathematics directly. Alluding to the trilogy of Gödel's theorems, uncertainty in quantum mechanics, and fractals,¹¹ Lyotard saw in contemporary mathematics

a current that calls into question precise measurement and prediction of the behavior of objects at the human scale ... postmodern science ... produces not the known, but the unknown.

Lyotard (1979)

Various authors have reminded readers that Gödel's theorems and the uncertainty principle (and chaos) are statements about formal systems in mathematics and particle physics (and nonlinear differential equations), respectively, and as such have no bearing on metaphysics.¹² The arguments are often eloquent but altogether beside the point, and of little comfort to seekers of certainty like Russell. Metaphysical certainty, whatever it may be, cannot be any less binding than a mathematical proof. Gödel's theorem, that it is impossible to prove, within a formal system, that that formal system is consistent, can reasonably be taken to mean that metaphysical certainty cannot be guaranteed by mathematical means alone.¹³ But Serre, in his comments to *Libération*, surely meant something more than the tautology that mathematical truth is totally reliable and

10. "Why don't you ask a physicist or a mathematician about difficulty?" was Derrida's response to a 1998 *New York Times* question about deconstruction: see Jacques Derrida, *Abstruse Theorist, dies at 74*, *New York Times*, October 10, 2004. Appeals to the presumed value of even the most abstruse mathematics, in order to legitimate obscurity elsewhere, are common. I first encountered such an argument in an article by composer (and former mathematician) Milton Babbitt entitled "Who cares if you listen?" (*High Fidelity*, February 1958): "Why should the layman be other than bored and puzzled by what he is unable to understand, music or anything else?" With this sort of talk, the justification of pure mathematics on aesthetic grounds is turned upside down. That is why I address aesthetic answers to the question of my title—which are by far the most popular among my colleagues—only in a footnote.

11. A cliché for the succeeding generation of literary critics: for a sample emphasizing chaos rather than Gödel, see N. Katherine Hayles (ed.), *Chaos and Order* (University of Chicago Press, 1991).

12. Much of *Prodiges et vertiges de l'analogie* by Jacques Bouveresse (*Raisons d'Agir*, 1999) is devoted to just this sort of reminder.

13. Predictably, religion steps in to fill the gap: see www.asa3.org/ASA/topics/Astronomy-Cosmology/PSCF9-89Hedman.html#16. John D. Barrow takes the implications of Gödel's theorems for physics seriously, while denying that they necessarily limit scientific objectivity (see, for example, "Domande senza risposta," in *Matematica e Cultura* 2002, edited by M. Emmer, pp. 13–24 (Springer, 2002)).

8. Otto Karnik, in "Attraction and repulsion," article in *Kai KeinRe-spekt*, p. 48, Exhibition Catalogue of the Institute of Contemporary Art (Bridge House Publishing, Boston, MA, 2004). The Hilbert quotation is easy to find but is probably apocryphal, which does not make it any less significant. *Mathematics and the Roots of Postmodern Thought*, by Vladimir Tasić, is an extended speculation on postmodernism's mathematical antecedents; see my review in *Notices of the American Mathematical Society* 50 (2003):790–99.

9. For example, "[Derrida's] thought is based on his disapproval of the search for an ultimate metaphysical certainty or source of meaning that has characterized most of Western philosophy." From the *Encyclopedia Britannica Online* (www.britannica.com).

verifiable *by the standards of mathematics*. The struggle to pin down this "something more," to find what one might call the "essence" of mathematics, is why the philosophy of mathematics keeps visiting the scenes of its many past defeats.

Even if Lyotard does not make the case very well, one can detect a "postmodern" sensibility in much of recent science, from Stephen Jay Gould's insistence that evolution is highly contingent, to complexity theory, to the study of consciousness as an "emergent" phenomenon. What these developments have in common is a rejection of reductionism and related top-down "master narratives," not because they are wrong but because they are irrelevant and useless. It would be going too far to describe this kind of science as a new Kuhnian paradigm (the notion is, in any case, widely criticized as oversimplified), but it is noticeably different from the disciplines that inspired the analytic philosophy of science. As for mathematics, there have been suggestions that it too has postmodern aspects—for example, Jürgen Jost has written a book entitled *Postmodern Analysis* and some specialists now claim to be working in "postmodern algebra"—but I do not see any genuine signs of this sensibility. Indeed, I am not even sure that it makes sense to draw the line between modern and postmodern. Hilbert's definition of mathematics as a game does sound like something from Derrida, but if Hilbert's foundational program ("wir müssen wissen, wir werden wissen") is not a prime example of high modernism, then what is? On the other hand, the abandonment of all forms of foundationalism in an anthology of Tymoczko (1998) is a rejection of "master narratives" within philosophy of mathematics, and indeed the blurb calls the anthology "postmodern."¹⁴

3 Sociology Aims for the High Ground

While Weil is supposed to have discounted Gödel's metaphysical menace by making it into a joke—"God exists since mathematics is consistent, and the Devil exists since we cannot prove it"—his fellow Bourbakist Dieudonné attempted a counterattack:

Just as physicists and biologists believe in the permanence of the laws of nature, solely because they have observed this up to now, ... the mathematicians called—wrongly—"formalists" (... at present the near totality of mathematical researchers) are convinced

that no contradiction will appear in set theory, none having manifested themselves for 80 years.¹⁵

This is either an inductive (empirical), sociological, or pragmatist argument. All these trends are indeed present in postmodernism, more typically in English sociology of science than in French philosophy:

The compelling force of mathematical procedures does not derive from their being transcendent, but from their being accepted and used by a group of people. The procedures are not used because they are correct, or correspond to an ideal; they are deemed correct because they are accepted.

David Bloor, in *Wittgenstein: A Social Theory of Knowledge* (Macmillan, London, 1983)

The Sociology of Scientific Knowledge (SSK) movement, of which David Bloor was a founder, is firmly rooted in postwar philosophy of science in the analytic tradition. The later Wittgenstein's discussion of mathematics, and knowledge more generally, in terms of "language-games," "forms of life," and learning to follow rules emphasizes social factors, and SSK is enthusiastically Wittgensteinian. Of course, Wittgenstein's work is notoriously unsystematic and lends itself to a variety of interpretations. I find it wrong to see the Wittgenstein who wrote "Grounds for *doubt* are lacking!" as a skeptic. Beyond the social factors to which Wittgenstein drew explicit attention, he clearly perceived "something more" specifically in mathematics ("the hardness of the logical must"), to which our language and philosophy are not able to do justice.¹⁶

Can sociology succeed where philosophy failed? Bloor's militant "naturalist" response to the question of "whether sociology can touch the very heart of mathematical knowledge" (Bloor 1976) is less an exercise in debunking metaphysics than an attempt to seize the metaphysical high ground for sociology. An otherwise subtle ethnographic study by Claude Rosental of the resolution of a conflict among logicians betrays a similar sensibility, as does his suggestion that training in mathematics and logic might have constituted a "serious handicap" to carrying out his project (Rosental

14. The anti-foundationalism of Tymoczko's anthology is largely inspired by Gödel's theorems.

15. Weil's joke is quoted in at least eighty-five sites found via Google; no primary source is given. Dieudonné's comment is naturally from *Pour l'Honneur de l'Esprit Humain*, pp. 244–45 (Hachette, 1987). Borel's remarks on the "self-correcting power of mathematics," in his contribution to the discussion of the article "Theoretical mathematics: toward a cultural synthesis of mathematics and theoretical physics" by A. Jaffe and F. Quinn, express a more modest form of pragmatism (*Bulletin of the American Mathematical Society* 29 (1993):1–13).

16. Quotations from Wittgenstein (1969, paragraph 4; 1958, paragraph 437).

2003). The classic declaration of the latter kind is due to Bruno Latour and Stephen Woolgar:

[W]e do *not* regard prior cognition... as a necessary prerequisite for understanding scientists' work. This is similar to an anthropologist's refusal to bow before the knowledge of a primitive sorcerer. There are, as far as we know, no a priori reasons for supposing that scientists' practice is more rational than that of outsiders.

Latour and Woolgar, *Laboratory Life*, pp. 29–30 (Princeton University Press, Princeton, NJ, 1986)

But one can also imagine sociologists paying serious attention to mathematicians' accounts of their experience, addressing in the process the question that Weil did not. For example, Bettina Heintz, in fieldwork at the Max-Planck-Institut in Bonn, which was billed as the first study of mathematics from the perspective of constructivist sociology of science, worries about "going native" and "overidentifying with the dominant culture." But her subject is the eminently sociological one of determining how mathematicians reach consensus, and her methodology, far from treating practicing mathematicians as "primitive sorcerers," records their epistemic perspectives sympathetically and at length. One has the impression that, in spite of the limitations of her methodology, Heintz is more interested in accounting for "real mathematics," to which we shall return below, whereas Bloor and Rosental are preoccupied with marshaling evidence to counter the metaphysical preoccupations of philosophers.

Under siege from Gödel's theorem, Popper's attack on verificationism, Kuhn's theory of scientific revolutions, Lakatos's dialectical approach to the contents of knowledge in *Proofs and Refutations*, as well as Wittgenstein, certainty in Russell's sense has largely been scrapped.¹⁷ As for the social, philosophical, and spiritual needs that the notion of metaphysical certainty was designed to address, they remain. Thus, on the one hand, those with tendencies that I have described as postmodernist continue to express skepticism regarding certainty, seemingly unaware that their

target is now little more than an advertising slogan that has little to do with the real concerns of mathematicians; while, on the other hand, analytic philosophy has sought to substitute more flexible notions. The term "warrant," for example, is used in an attempt by Philip Kitcher to develop a consistent account of mathematics on empirical rather than aprioristic grounds. Kitcher recalls FREGE's [VI.56] frustration with the mathematicians of his time, observing that, "When Frege emphasizes the possibility of complete clarity and certainty in mathematical knowledge, he is advancing a picture of mathematics that is almost irrelevant to the working mathematician" (Kitcher 1984). However, Kitcher and the SSK remain obsessed by the problem of "how our mathematical knowledge [is] acquired" (Kitcher 1984), where knowledge is taken to be true and justified belief.

Reading Heintz (2000), one learns that now, as in Frege's day, mathematicians themselves widely consider these problems to be outdated or beside the point. The most controversial aspect of the SSK's "strong programme," formulated by Bloor and Barry Barnes, is the "thesis of symmetry": the insistence that truth or falsity not be taken into account when investigating how a scientific claim comes to be accepted as knowledge. Heintz's fieldwork suggests that this is compatible with the view prevailing among mathematicians regarding acceptance of a mathematical proof, a "kind of consensus theory of truth" (Heintz 2000).¹⁸

A striking instance of "how a mathematical proof comes to be accepted as knowledge" is playing out even as I am writing these lines. Grigori Perelman's announced proof of THE POINCARÉ CONJECTURE [V.28] is undergoing unprecedented scrutiny in a small number of specialized centers, with the hope of determining the truth or falsity of Perelman's claim. This is going on completely beyond the spotlight of sociology, as far as I know, and with no guidance from philosophy, even though the \$1 000 000 prize offered by the Clay Mathematical Institute is in no sense Pla-

17. Lakatos's posthumous "A renaissance of empiricism in the recent philosophy of mathematics," presents a long series of quotations by mathematicians and a few philosophers, including Russell in 1924, acknowledging that mathematics is uncertain, after all. Naturally, most of those cited refer directly or indirectly to Gödel's theorem. The article was reprinted in Tymoczko (1998). "Only dogma or theory has made people say that mathematics as a whole has a peculiar certainty," writes Hacking (2000). Certainty persists, however, in the titles of philosophy books, e.g., Marcus Giaquinto's optimistic *The Search for Certainty: A Philosophical Account of Foundations of Mathematics* (Oxford University Press, Oxford, 2004).

18. Heintz quotes Yu. I. Manin—"A proof only becomes a proof after the social act of 'accepting it as a proof'"—as well as René Thom's "community" theory of truth. One can of course always ask whether Heintz selectively quoted mathematicians whose positions support her thesis. This question can be asked of any sociological study, and it is best to let the sociologists work out their methodological issues. An important remark, however: though Heintz's original goal was to account for the formation of consensus among mathematicians within a science studies framework—with questionable success, but that is another matter—she does not defend a particular school within philosophy of mathematics. In this she differs from Bloor, for instance, who identifies himself explicitly as an empiricist.

tonic,¹⁹ and the rules for awarding the prize presuppose the fallibility of the mathematical community, in terms very similar to those that Heintz's informants expressed spontaneously (see the third and subsequent paragraphs at www.claymath.org/millennium/rules_etc). The case is exceptional, however; "certifying knowledge," in Rosental's sense, is as such relatively unimportant to mathematicians, and Perelman's close readers would probably describe what they are doing as attempting to *understand* his proof rather than "certifying" it as knowledge (for the sake of the community, or a generous benefactor, or philosophers or sociologists).²⁰

4 Truth and Knowledge

"By far the larger part of activity in what goes by the name *philosophy of mathematics* is dead to what mathematicians think and have thought, aside from an unbalanced interest in the 'foundational' ideas of the 1880–1930 period, yielding too often a distorted picture of that time," announced David Corfield, presenting his efforts to develop a "philosophy of real mathematics."²¹ Corfield contrasted the traditional apriorist's concerns, "How should we talk about mathematical truth? Do mathematical terms or statements refer? If so, what are the referents and how do we have access to them?" (Corfield 2003), with the list of questions Aspray and Kitcher consider typical of the "maverick tradition" in philosophy of mathematics: "How does mathematical knowledge grow? What is mathematical progress? What makes some mathematical ideas (or theories) better than others? What is mathematical explanation?" (quoted by Corfield 2003).

The mavericks, well represented in Tymoczko's anthology, have moved a welcome step away from cer-

tainty. Nevertheless, the philosophers and philosophically minded sociologists I have mentioned—with the partial exception of Corfield, to be explored below—still often write as though mathematicians were creating Truth or Knowledge,²² almost as a favor to philosophy or sociology, to show how such a feat is possible. Or just to show that it *is* possible.²³ We mathematicians, on the other hand, are quite convinced that we are creating *mathematics*, and it is the "why" of that activity, without the ennobling assimilation to the generic objects of interest to epistemology, that, as Weil understood, required no explanation in Helsinki.

"Whoever undertakes to set himself up as a judge in the field of Truth and Knowledge is shipwrecked by the laughter of the gods," wrote Einstein. Mathematicians tend to respond with dismay rather than laughter, and then only to blunders so egregious as to be universally recognized as such.²⁴ Although those who find fault with philosophical speculation regarding the nature of mathematics seem to be under an implicit obligation to propose a speculative alternative, experience suggests that the practice of mathematics renders one unfit to do so. This, more than the fear of ridicule, is the main reason I would not venture my own speculative philosophy of mathematics. If it is hard "for those who are used to thought processes stemming from geometry and algebra" to "develop the sort of intuition common among physicists" (R. MacPherson, quoted in *Quantum Fields and Strings: A Course for Mathematicians* by Deligne (volume 1, p. 2)), bridging the gap between mathematicians and metaphysicians is probably hopeless. There are superficial parallels, to be sure: a metaphysical abstraction such as "essence," like a mathematical abstraction such as "set," designates nothing in itself, but rather refers to a canonical body of special-

19. This article was written in late 2004. The proof is now accepted as correct, and in 2006 Perelman was offered a Fields medal, which he declined. He has also refused the Clay Mathematical Institute prize.

20. "Having shown how the production of certified knowledge in logic could constitute an object of a sociological investigation and analysis, a vast field of research takes shape" (Rosental 2003). I suspect that identifying and accounting for the priorities expressed by mathematicians themselves would constitute a much richer field of research.

21. The quotation is from Corfield's *Towards a Philosophy of Real Mathematics* (Corfield 2003). Compare it with Ian Hacking's comment that "the most striking single feature of [twentieth century philosophy of mathematics] is that it is very largely banal" (Hacking 2002). For Hacking's philosophy of mathematics, see his *What Mathematics Has Done*.

"Real mathematics," for Corfield, who is remarkably well-informed about trends in the most diverse branches of mathematics, is "real" in the same way as "real ale." I readily agree that skepticism toward this sort of realism is self-defeating.

22. See, however, Hacking: "The truth of a sentence (of a kind introduced by a style of reasoning) is what we find out by reasoning using that style" (Hacking 2002).

23. Many of the authors in Tymoczko (1998) also look to the (real) practice of mathematics for philosophical insight, but Truth and Knowledge keep creeping in. Arriving in France in 1994, I was astonished to discover that the concerns of twentieth-century French philosophers of mathematics are entirely different. Following Husserl, the French concentrate largely on the phenomenological experience of the individual mathematical subject. It is only a slight exaggeration to say that the French-language and English-language traditions in philosophy of mathematics have become mutually incomprehensible. Fortunately, mathematicians writing in French and in English have no trouble citing each others' works.

24. As Serre put it in his comments to *Libération*, "Si vous ne voulez pas que les choses soient parfaites, ne faites pas de maths." Heintz's book is an inquiry into the roots of this apparent universal tendency to consensus, and finds it in the institution of the proof; Rosental treats a (highly unusual) case in which universal consensus apparently failed. The Einstein quotation is in Kline (1980).

ized texts in which the term plays a central role. I would like to argue that the nothing designated by “set” is somehow different, and more fruitful, than the nothing designated by “essence.” But the means at my disposal for making such an argument take the form of mathematical reasoning, which leads me, at best, to a vicious circle.²⁵ More bluntly, and for reasons akin to those Serre invoked in his *Libération* interview, I cannot be satisfied with an answer that is less certain than the sort of answer mathematics provides; for a mathematician, a pragmatist answer to Weil’s question is an admission of defeat. And yet I am aware that (metaphysically certain) grounds for distinguishing mathematical certainty from pragmatic certainty are lacking!

Another, possibly more profound, reason to steer clear of speculation is that, whereas philosophy presents itself as a dialogue extending over millennia, so that to understand each new contribution one would ideally be familiar with all previous contributions, mathematics is in principle supposed to be derivable by pure reason from a small number of axioms. A philosophical proposition, in other words, remains attached to its origins and context; a mathematical proposition floats free. This principle, an important constituent of the aura of metaphysical certainty surrounding mathematics, does not in fact bear much resemblance to mathematics as it is actually practiced—“one of humankind’s longest conversations,” as Barry Mazur puts it. I am nonetheless painfully aware that my personal “conversation” with the philosophical tradition is thoroughly unreliable, and my choice of footnotes is primarily the fruit of a random walk (or random surf, or remix) among scraps of the literature I have happened to encounter.

If I am nevertheless writing about philosophy, it is in large part because of a question that was put to me in 1995, during a presentation of Wiles’s proof of FERMAT’S LAST THEOREM [V.12] to an audience of scientists. An October 1993 article in *Scientific American* entitled “The death of proof” had called Wiles’s proof a “splendid anachronism,” citing Laszlo Babai and his collaborators, among others, in support of the thesis that, in the future, deductive proof in mathematics will be largely supplanted by computer-assisted proofs and probabilistic arguments. That same month the *Notices of the American Mathematical Society* (40:

978–81) published Doron Zeilberger’s manifesto “Theorems for a price,” predicting a rapid transition from rigorous proofs to an “age of *semi-rigorous* mathematics, in which identities (and perhaps other kinds of theorems) will carry price tags” measured in computer time and proportional to the degree of certainty desired, to be followed in turn by “abandoning the task of keeping track of price altogether, and ... the metamorphosis to non-rigorous mathematics” (John Horgan, *Scientific American* October 1993:92–102).²⁶

Feeling called upon to answer the question Weil avoided, I argued that the basic unit of mathematics is the concept rather than the theorem, that the purpose of a proof is to illuminate a concept rather than merely confirm a theorem, and that the replacement of deductive proofs by probabilistic or mechanical proofs should be compared not to the introduction of a new technology for producing shoes, say, but rather to the attempt to replace shoes by the sales receipts, or perhaps the cash profits, of the shoe factory. The audience had its own question: Was I talking about certainty? Of course not. That option has been philosophically discredited, as I have tried to explain. And other normative prescriptions fall victim no less easily to the laughter of philosophers. On the other hand, I see no pragmatic reason why probabilistic or mechanical proofs would not suit the five goals on the AMS committee’s list just as well as deductive proofs, nor any sociological reason why they should not be as effective in commanding consensus in the event of a paradigm shift. So what was I talking about?

Such a question, at this point in the essay, practically begs to be answered by an advertising slogan. For example:

The practice of making what one writes “reliable and verifiable” fosters critical thinking in general.

This is a popular argument for teaching proofs, and is probably even true, but how would one go about verifying such a claim? I am very much tempted to say that the concepts that serve as material for “one of humankind’s longest conversations,” deserve to be

25. “Truth is always the possibility of its proper destruction,” according to the (nonpostmodern) French philosopher, Alain Badiou, taking Gödel’s theorem as an example (www.egs.edu/faculty/badiou/badiou-truth-process-2002.html).

26. In the pop posthumanist scenarios promoted by Hans Moravec, Ray Kurzweil, and the like, computers acquire all human capabilities, including the ability to generate and prove theorems—for some reason this is always considered a landmark—by the middle of the twenty-first century. The distinction between humans and computers subsequently fades away rather rapidly, making Zeilberger’s prediction moot.

A more recent, and much more nuanced, discussion of prospects for automatic theorem proving has been posted on the Internet by Maggesi and Simpson (undated).

appreciated on their own terms. Note that nothing is more "emergent" than a conversation. But that would be unfaithful to the spirit of Mazur's book, one of whose strengths is its refusal to conform to a linear narrative. In any case, on its own the argument does not seem to be sufficient: a similar argument could be made in favor of religious faith.

5 "Ideas, Even Dreams"

Rather than hazard an answer to Weil's (non)question here, I will take a cue from Corfield and suggest that one can best account for the value of pure mathematics by attending to what mathematicians write and say. A handful of commonplace words appear consistently, invested with unexpected power, when mathematicians attempt to account, formally as well as informally, for their value judgments, and these collectively constitute an answer to the question Weil left hanging.

WEYL [VI.80] wrote a book with the provocative title *The Idea of a Riemann Surface*²⁷ and referred in his preface to Plato. The word "concept" that was central in my reply to the audience is closer to this use of the term "Idea" as used by any number of philosophers, including most of those mentioned in this essay. A square, or a RIEMANNIAN MANIFOLD [I.3 §6.10], would be a concept or "Idea" in this sense, and this is how the word "concept" tends to be used by mathematicians, who generally reserve the word "idea" to designate something else. In Plato's *Meno*, the proof of the doubling of the square—draw diagonals and fit the resulting triangles together—which the slave "remembers" under Socrates' coaching, is taken by Plato to be contained in the "Idea" of the square. For a mathematician, drawing the diagonals and moving the triangles *are* the ideas.

That a contrast can be drawn, as I did in 1995, between "illuminating concepts" and "confirming theorems" is something of a truism among mathematicians and even some philosophers. Even by 1950 Popper had argued that "a calculator ... will not distinguish ingenious proofs and interesting theorems from dull and uninteresting ones" (quoted in Heintz 2000). Corfield correctly states that "what mathematicians are largely looking for from each other's proofs are new concepts, techniques, and interpretations"; they are not merely "establishing the truth or correctness of propositions" (p. 56). However, although he devotes a chap-

ter to the "extremely complex subject" of "mathematical conceptualization," he does not dwell on concepts (or "Ideas") as such; and neither will I. It is almost impossible to talk in general terms about mathematical concepts without getting caught up in the debate over their reality (and provoking the laughter of the philosophers). Those who write about mathematics (mathematicians included: see Hersh (1997)) have an irritating tendency to claim that most mathematicians are Platonists, whether or not they have committed themselves explicitly to a philosophical position. Maybe it can be argued that Platonism is implicit in the syntax of mathematical statements; maybe this is what Weil meant by his claim, quoted by Bourguignon (2001), that most mathematicians "spend a good portion of their professional time behaving as if they were [Platonists]." ²⁸ In practice I would guess that most mathematicians are pragmatists, in the spirit of the remarks of Dieudonné quoted above.

On the other hand, there is no doubt whatsoever that the "ideas" that matter to mathematicians are real. A mathematician, according to a joke attributed to Weil,²⁹ can be defined as someone who has had two ideas (mathematical, of course). But then, Weil worried, so-and-so would be a mathematician. In a celebrated account by POINCARÉ [VI.61] of the role of the unconscious in a mathematical discovery, the climactic moment was the arrival of an idea ("the idea came to me") as he placed his foot on the steps of the omnibus ("L'idée me vint," Poincaré (1999)).

More to the point, consider Hacking's justification of his own commitment to a realist ontology of electrons: "*So far as I'm concerned, if you can spray them then they are real*" (Hacking 1983). By the same token, if you can steal ideas, then they are real. Every mathematician knows that ideas can be and often are stolen. Polemics then ensue, considerably juicier than the epistemic controversy studied by Rosental.

Nothing in the life of mathematics has more of the attributes of materiality than (lowercase) ideas. They have "features" (Gowers 2002), they can be "tried out" (Singer³⁰), they can be "passed from hand to hand" (Corfield 2003), they sometimes "originate in the real

27. Weyl used the word *Idee* in his title but applied the term *Begriff* (concept) elsewhere in the text. Both terms arrived in English as "concept."

28. Plato saw things quite the other way around: "Their language [speaking of mathematicians] is most ludicrous, *though they cannot help it*, for they speak as if they were doing something and as if all their words were directed toward action" (*Republic* VII.527a, my emphasis).

29. I heard this joke reported by several people who claimed to have heard it from Shimura, and I believe but am not certain that I too first heard it from Shimura.

30. Quoted at www.abelprisen.no/en/prisvinnere/2004/interview_2004_7.html.

world" (Atiyah in the preface to Arnold et al. (2000)) or are promoted from the status of calculations by becoming "an integral part of the theory" (Godement 2001). At some point they come into being: it is generally understood, for example, that "new ideas" will be needed to solve the Clay Millennium Problems. They can also be counted. I once heard Serre introduce the proof of a famous conjecture by saying that it contained two or three real ideas, where "real" was intended as high praise. The ambiguity did not concern the number of ideas—there were three, which Serre enumerated—but whether all three were original to the author. Ideas are public: necessarily so, in order to be stolen, or to be presentable as Serre did in his lecture. Poincaré's idea was a sentence ("the transformations of which I had made use to define Fuchsian functions were identical to those of non-Euclidean geometry"); the slave's idea in *Meno* was a line in the sand.

Early in his unpublished memoirs *Récoltes et Semailles*, Grothendieck wrote that "ideas, even dreams" were, in Allyn Jackson's terminology, the "essence and power" of his mathematical work (Jackson 2004). An idea is typically symptomatic of "insight," and the capacity for insight is generally called "intuition." Mathematicians have borrowed all of these terms from philosophy but use them to completely different ends. Philosophers tend to follow Kant in attributing intuitions—the ones that without concepts are blind—to transcendental subjects or their more down-to-earth offsprings. Intuition in this sense is a poor substitute for certainty, as even the mavericks recognize. "Intuition ... is frequently a *prelude* to mathematical knowledge," wrote Kitcher. "By itself it does not warrant belief." Poincaré called intuition "the tool of invention," a "je ne sais quoi" that holds a proof together, but he contrasted it with logic, "the tool of demonstration," which "alone can provide certainty." Saunders Mac Lane expressed himself in much the same terms nearly a century later. David Ruelle considered reliance on (visual) intuition a characteristic feature of human (as opposed to extraterrestrial) mathematics.³¹

In each case intuition belongs to the *private* sphere, and is relegated to the "context of discovery," as opposed to the "context of justification" deemed worthy of philosophy's full attention. When mathematicians refer to "intuition" in the sense I have in mind, it

is crucially *public*.³² As in the quotation from MacPherson a few paragraphs back, it can be transmitted from teacher to student, or through a successful lecture, or developed collectively by running a seminar and writing a book on the proceedings. It has something in common with a "style of reasoning," but on a smaller scale. Grothendieck resorted to perceptual metaphor when describing Serre's ability to communicate something akin to intuition:

The essential thing was that Serre each time strongly sensed the rich meaning behind a statement that, on the page, would doubtless have left me neither hot nor cold—and that he could "transmit" this perception of a rich, tangible, and mysterious substance—this perception that is at the same time the *desire* to understand this substance, to penetrate it.

Récoltes et Semailles, p. 556

"Even those who try to articulate, to classify, the fruits of the imagination, and who are committed to the existence of an inner experience concomitant with it, admit to dark difficulty in describing it," wrote Mazur, elaborating an unusual array of literary and rhetorical strategies to chip away at the difficulty (Mazur 2003). This much is certain: this inner experience of imagination, or of understanding, is what drives people to become mathematicians, and it is why Weil could count on his audience's silent assent. Heintz recorded some of her informants' attempts to describe this inner experience: "[In mathematics] you have concrete objects before you and you interact with them, talk with them. And sometimes they answer you." She even talks about the "idea" that helps put the pieces together. "And suddenly you see the picture," she was told. Yet all this raw ethnographic data is presented in a chapter whose title, "Beauty and experiment: discovery of truth in mathematics," betrays her relentlessly epistemological preoccupations (Heintz 2000).

"The specific ways that mathematical truths move from person to person, and how they are transformed in the process, are as difficult to capture as the truths themselves," wrote Mazur (2003), in what could have been a comment on Grothendieck's remarks on Serre. The central notion in Mazur's book is that of "imagination." I have chosen the terms "idea" and "intuition" not for their intrinsic importance, though I believe each of the terms points to ways of talking about the famous

31. Kitcher (1984, p. 61); Poincaré (1970, pp. 36–37); Mac Lane, in his contribution to the discussion of the Jaffe–Quinn article cited in note 15, *Bulletin of the American Mathematical Society* 30 (1994): 178–207; Ruelle's quote is from an article entitled "Conversations on mathematics with a visitor from outer space" from Arnold et al. (2000).

32. This is also true of the normative program of intuitionism associated with BROUWER [VL75], but that is definitely not what I have in mind.

"flash in the middle of a long night" that ends Poincaré's *The Value of Science*: "But this flash is everything." What strikes me about these terms is how their pervasiveness in mathematicians' conversations—the sense that they, more than the definitive theorems, are "everything"—contrasts so starkly with their near exclusion from philosophical consideration, even though the words themselves can be seen on practically every page of philosophy of mathematics. Maybe their very banality makes them appear philosophically trivial. Or maybe the problem is that the same words serve so many distinct purposes. Corfield uses the same word to designate what I am calling "ideas" ("the ideas in Hopf's 1942 paper") as well as "Ideas" ("the idea of groups") and something halfway between the two (the "idea" of decomposing representations into their irreducible components for a variety of purposes, p. 206). Elsewhere, the word crops up in connection with what mathematicians often call "philosophy," as in the "Langlands philosophy" ("Kronecker's ideas" about divisibility, p. 202), and in many completely unrelated places as well. Corfield proposes to resolve what he sees as an anomaly in Lakatos's "methodology of scientific research programmes" as applied to mathematics by "a shift of perspective from seeing a mathematical theory as a collection of statements making truth claims, to seeing it as the clarification and elaboration of certain central ideas" (p. 181). He sees "a kind of creative vagueness to the central idea" in each of the four examples he offers to represent this shift of perspective; but on my count the ideas he chooses include two "philosophies," one "Idea," and one which is neither of these.

Other value-laden terms are no less important. In the wake of BOURBAKI [VI.96], quite a few philosophers (Cavaillès, Lautbaki, Piaget, and more recently Tiles) have made serious attempts to make sense of "structure" in mathematics. I have read a number of philosophical attempts to account for mathematical aesthetics, though none has left much of an impression. The practically universal use of dynamical or spatiotemporal metaphors ("the space X is fibered over Y ," etc.), and the pronounced tendency to present proofs as series of actions playing out in time ("now choose an orbit passing arbitrarily close to the point x ") have attracted little attention from philosophers.³³ These phenomena

may be linked to the curious preference of many mathematicians for blackboards over contemporary audio-visual technology, which in turn draws attention to the neglected (and emergent) aspect of mathematical communication as *performance*, a word that manages to be typically postmodern and premodern at the same time.

For his part, Corfield does not talk much about "intuition" and is ambiguous about what he means by "ideas," but his discussions of "natural" and "importance," in the context of an analysis of the debate on the relative merits of groups and groupoids, are philosophically insightful while remaining faithful to the use of the terms by "real" mathematicians. His treatment of "postmodern algebra," where "diagrams are not just there to illustrate, they are used to calculate and to prove results rigorously" (p. 254), also has street credibility. It is true that much of his book remains concerned with "maverick" questions, such as accounting for plausible reasoning. But there is no question that Corfield likes mathematics, and for the right reasons; his book, unlike most treatises in philosophy of mathematics, is definitely part of the "conversation."

Morris Kline called the "loss of certainty" entailed by Gödel's theorems an "intellectual tragedy" and actually counseled "prudence" in designing bridges "using theory involving infinite sets or the axiom of choice" (Kline 1980). The word "tragedy" seems misplaced but the pathos is real, as it was for Russell. Pathos and its twin, resolute optimism, have found an unlikely home in the philosophy of mathematics:

If this conception of mathematics [as "human knowledge of structures gained by employing reason beyond the bounds of logic"] can be sustained, mathematics could once again serve as a source of an image of reason liberated from formal imprisonment, freed to confront apocalyptic post-modern visions.

Mary Tiles, *Mathematics and the Image of Reason*, p. 4
(Routledge, London, 1991)

Whether or not it carries weight with congressional committees, I find this goal appealing, but it is a goal for philosophers, not for mathematicians. I'm willing to apply the "principle of charity" to philosophers if they will do the same for me. Corfield wrote (p. 39):

Human mathematicians pride themselves on producing beautiful, clear, explanatory proofs, and devote much of their effort to reworking results in conceptually illuminating ways. Philosophers must not evade their duty to treat these value judgments in mathematics.

33. Nuñez's article "Do *real* numbers really move?" (in Hersh 2006) makes interesting points regarding mathematicians' use of metaphors of motion, though he limits his analysis to examples specifically related to the mathematics of motion. Plato specifically disapproved of mathematicians' use of action metaphors.

They also have a duty, it seems to me, to account for terms like “idea” and “intuition”—and “conceptual” for that matter. An answer to the question “Why philosophy?” might well begin there.

Postscript

In December 2004 my university joined a number of other institutions in France and elsewhere in hosting a traveling UNESCO-sponsored exhibition entitled “Pourquoi les mathématiques?” Hoping to learn the answer before my submission deadline, I spent a few hours at the exhibition. It was clever and engaging, presenting a variety of (pure) mathematical ideas with a sprinkling of practical applications, but in no way did it address the “Pourquoi?” of the title. An organizer was on hand, and when I turned to her for guidance she explained that the French title was a solution to a problem of translation. The English title, which came first, was “Experiencing mathematics.” This, she assured me, has no adequate French translation, so “Pourquoi les mathématiques?” was chosen as the best substitute.

Maybe the solution to the problem of my title is simply to accept the translation in the opposite direction. Even the most ruthless funding agency is not yet so post-human as to require an answer to “Why experience?”³⁴

Acknowledgments. I thank Cathérine Goldstein and Norbert Schappacher for pointing me in the directions of the Rosental and Heintz books, among other source material, and for vigorously criticizing my project as well as its execution. I also thank Mireille Chaleyat-Maurel for explaining the title of the UNESCO exhibition and Ian Hacking for critically reading an earlier version of the manuscript with tolerance and rigor. David Corfield receives thanks for several helpful clarifications. Barry Mazur is thanked especially warmly for many suggestions, much encouragement, for help with the title, and most of all for showing, in his *Imagining Numbers*, that there is at least one way out of the fly-bottle.

Further Reading

Arnold, V., et al. 2000. *Mathematics: Frontiers and Perspectives*. Providence, RI: American Mathematical Society.
 Barthes, R. 1967. *Système de la Mode*. Paris: Éditions du Seuil.
 Bloor, D. 1976. *Knowledge and Social Imagery*. Chicago, IL: University of Chicago Press.

Bourguignon, J.-P. 2001. A basis for a new relationship between mathematics and society. In *Mathematics Unlimited—2001 and Beyond*, edited by B. Engquist and W. Schmid. New York: Springer.
 Corfield, D. 2003. *Towards a Philosophy of Real Mathematics*. Oxford: Oxford University Press.
 Godement, R. 2001. *Analyse Mathématique I*. New York: Springer.
 Gowers, W. T. 2002. *Mathematics: A Very Short Introduction*. Oxford: Oxford University Press.
 Hacking, I. 1983. *Representing and Intervening*. Cambridge: Cambridge University Press.
 ——. 2000. What mathematics has done to some and only some philosophers. *Proceedings of the British Academy* 103:83–138.
 ——. 2002. *Historical Ontology*. Cambridge, MA: Harvard University Press.
 Harvey, D. 1989. *The Condition of Postmodernity*. Oxford: Basil Blackwell.
 Heintz, B. 2000. *Die Innenwelt der Mathematik*. New York: Springer.
 Hersh, R. 1997. *What Is Mathematics, Really?* Oxford: Oxford University Press.
 ——, ed. 2006. *18 Unconventional Essays on the Nature of Mathematics*. New York: Springer.
 Jackson, A. 2004. Comme appelé du néant—as if summoned from the void: the life of Alexandre Grothendieck. *Notices of the American Mathematical Society* 51:1038.
 Kitcher, P. 1984. *The Nature of Mathematical Knowledge*. Oxford: Oxford University Press.
 Kline, M. 1980. *Mathematics: The Loss of Certainty*. Oxford: Oxford University Press.
 Lakoff, G., and R. E. Núñez. 2000. *Where Mathematics Comes From*. New York: Basic Books.
 Lloyd, G.E.R. 2002. *The Ambitions of Curiosity*, p. 137, note 13. Cambridge: Cambridge University Press.
 Lyotard, J.-F. 1979. *La Condition Postmoderne*. Paris: Minuit.
 Maggesi, M., and C. Simpson. Undated. Information technology implications for mathematics, a view from the French Riviera. (Available at <http://math1.unice.fr/~carlos/preprints/gzloc/arxiv/gzexp.ps> (apparently not posted before 2004).)
 Mancosu, P., ed. 1998. The current epistemological situation in mathematics. In *From Brouwer to Hilbert. The Debate on the Foundations of Mathematics in the 1920s*. Oxford: Oxford University Press.
 Mazur, B. 2003. *Imagining Numbers (Particularly the Square Root of Minus Fifteen)*. New York: Farrar Straus Giroux.
 Minsky, M. 1985/1986. *The Society of Mind*. New York: Simon and Schuster.
 Poincaré, H. 1970. *La Valeur de la Science*. Paris: Flammarion.
 ——. 1999. *Science et méthode*. Paris: Éditions Kimé.
 Rosental, C. 2003. *La Trame de l'Évidence*. Paris: Presses Universitaires de France.

34. Or, as Weyl put it, “with [mathematics] we stand precisely at the point of intersection of restraint and freedom that makes up the essence of man itself.” Note the word “essence” (see Mancosu 1998). I thank David Corfield for this quotation.

- Tymoczko, T., ed. 1998. *New Directions in the Philosophy of Mathematics*. Princeton, NJ: Princeton University Press. (First published in 1986.)
- Wittgenstein, L. 1958. *Philosophical Investigations*, volume I. Oxford: Basil Blackwell.
- . 1969. *On Certainty*. Oxford: Basil Blackwell.

VIII.3 The Ubiquity of Mathematics

T. W. Körner

1 Introduction

We live surrounded by mathematics: when we open a door or use a nutcracker, we exploit ARCHIMEDES' [??] law of the lever; when a bus goes around the corner, we experience at first hand NEWTON's [VI.14] law that a body continues to travel in uniform motion in a straight line unless acted on by an external force; when we use a rapidly accelerating elevator, we can feel for ourselves the equivalence of gravitational and accelerational inertia that lies at the heart of GENERAL RELATIVITY [IV.13]; and when we run a tap fast into a kitchen sink we see a thin and flat circle of water with a clear boundary, which is the chaotic "hydraulic jump" between two well-behaved solutions of a certain PARTIAL DIFFERENTIAL EQUATION [I.3 §5.4].

Because mathematics and physics are so interlinked, almost everything we see involves mathematics. With the help of elementary calculus, we know that a baseball, after it leaves the bat, will have a trajectory in the shape of a parabola. This calculation assumes that there is no air resistance, but a more complicated calculation can take air resistance into account too. If a chain hangs between two points, then the curve it forms can again be explained mathematically. This time, the technique used is the CALCULUS OF VARIATIONS [III.96]: the curve is the one that minimizes the potential energy of the chain, and the calculus of variations allows you to work it out. (It is called a *catenary*. The rough idea of the calculation is to consider small perturbations of the chain. Since the potential energy is minimized, we know that however we perturb it, we cannot decrease the potential energy. This information can be used to derive a differential equation that determines the curve. In general, the differential equations that arise from this technique are called the *Euler-Lagrange equations*.) Even the way that wet sand behaves when you walk across it involves interesting mathematics, as Reynolds realized in 1885. Typically, the sand just around where you tread dries out—if you

have not noticed this strange phenomenon, then have a look next time you are on a beach. The reason this occurs is that when the tide goes out the sea tends to leave the grains of sand extremely well-packed. If you then tread on the sand, you disturb this packing, creating a less well-packed part of the sand near where you tread. This has more room for water, so it draws water in and down, temporarily drying out the sand around your foot.

It would be easy to give hundreds more examples of physical phenomena that can be analyzed mathematically. However, if one accepts that physics governs the universe and that mathematics is the language of physics, then it is not surprising that these applications exist. Therefore, this article will focus on the appearance of mathematics in other areas, and in particular geography, design, biology, communication, and sociology.

2 Uses of Geometry

If you travel about on Earth's surface, then you need to make small adjustments to your watch as you move from one time zone to another. There is one exception to this, however: if you cross the international date line, then you have to make a big adjustment (assuming, that is, that your watch shows not just the time but the date as well). Why is it necessary to have a discontinuity of this kind? Well, suppose that it is midnight on a Tuesday in Lisbon, for example, and imagine a path that goes westward right around the globe. If the time changes along this path are all small ones that reflect where one is in relation to the sun, then the time of day goes back by one hour for every 15 degrees of longitude that we move. Therefore, when one gets back to Lisbon it is midnight on Monday. (Remember that we are talking about a *mental* path here, and not an actual journey.) Something is clearly not right. The practical consequences of this theoretical problem were first felt by the tattered remnants of Magellan's first circumnavigation of the globe who had to do penance for performing religious ceremonies on the wrong day!

Here is another argument for the necessity of the date line. Let us ask exactly when the year 2000 began. The answer depends, of course, on what part of the world you are talking about, and more particularly on its longitude, but for any part the answer is midnight at the beginning of January 1. In other words, in any particular place the year began when the Sun was (approximately) over the opposite side of the world. It follows

that at any given time at most a small fraction of the world was celebrating the very beginning of the year 2000. Therefore, at least somewhere had to go first, which means that parts of the world just to the east of it had missed their chance and had to wait almost 24 hours. Thus, again we see that there has to be a discontinuity.

These phenomena reflect the fact that a certain continuous map has no continuous inverse. The map in question takes a real number w to the point $w \mapsto (\cos w, \sin w)$, which lives in the unit circle. Notice that if we add 2π to w then we do not affect the values of $\cos w$ and $\sin w$. Now let us try to invert the map. This means that for each point (x, y) in the unit circle we must pick some w such that $\cos w = x$ and $\sin w = y$. This w is the angle that the line from 0 to (x, y) makes with the horizontal, with the all-important proviso that you can add any multiple of 2π to it. So the question becomes, can we choose the appropriate multiple in a continuous way? Again, the answer is no, since if you go around the circle once and let the angle vary continuously, you find that you have added 2π to it.

The above fact is one of the simplest theorems of TOPOLOGY [IV.6], the branch of mathematics that you turn to if you want to know about the existence or non-existence of continuous functions with given properties. Another situation where continuous functions are useful is when one is creating a map (in the geographer's sense) of the world. Such maps are more convenient if they are drawn on a flat piece of paper, so a preliminary question we might ask is whether there is a continuous function from the surface of the sphere to the plane such that any two different points in the sphere go to different points in the plane. Not only is the answer no, but *Borsuk's antipodal theorem* tells us that there must be some pair of *antipodal* points (that is, points of the sphere that are exactly opposite each other, such as the North and South Poles) that go to the same point in the plane.

However, perhaps we do not mind too much about continuity. If we take our sphere and make a cut from the North Pole to the South Pole, then we can open it up at the cut and flatten it out onto a plane. (To see this, imagine that it is made of particularly stretchy rubber.) Alternatively, we could cut the sphere into two hemispheres and draw maps of each hemisphere separately.

Now another problem arises: it does not seem to be possible to draw a map of even half the world without distortions. This is not a topological problem, but

a *geometrical* one, in the sense that we are interested in finer properties of Earth's surface—shape, angle, area, and so on—than those that are preserved by continuity. Because the sphere has positive CURVATURE [III.80], no part of it can be mapped to the plane in a length-preserving manner, so some distortion is necessary. However, we have a certain amount of freedom to decide what kind of distortion we are prepared to accept and what kind we would like to avoid. There is, it turns out, a *conformal map* from the sphere (minus the poles) to a cylinder (which one can cut and roll out so that it fits into a plane)—it is the famous “Mercator projection.” A conformal map is one that preserves angles, so the Mercator projection is particularly useful for navigation purposes: if it looks as though you need to head north-northwest, then you really do. A disadvantage of the Mercator projection is that as you move away from the equator, the countries look bigger and bigger (though the angle-preserving property means that in close-up they are always the right shape). There is another projection that distorts shapes but preserves area. To work out the details of these projections, one must use mathematics—and in particular solve differential equations.

Here are a few simple applications of geometry to everyday life. If you have ever wondered what the best shape is for a manhole cover, then mathematics can come to your aid. Of course, it depends what one means by “best,” but if you often need to lift manhole covers, then you may be annoyed if they keep falling down the manholes. Can this be avoided? If the cover is rectangular, then the length of any side is less than the length of the diagonal, so it can drop down the hole, but if it is circular, then its width is the same in all directions and this is not possible.

Does this mean that only circular manhole covers are safe from dropping down their manholes? Actually, no. If you draw the three vertices of an equilateral triangle and join each pair of them by a circular arc centered on the third, then you obtain a sort of “curved triangle,” known as the *Reuleaux triangle*. (This is commonly misspelt “Rouleaux” in the mistaken belief that it has something to do with rolling. Actually, it is named after a nineteenth-century German engineer called Franz Reuleaux.)

Have you ever wondered why coins are the shapes they are? Most of them are circular, but the British fifty pence piece, for example, is a slightly curved polygon with seven sides. A moment's thought makes it clear that for any odd number $n \geq 3$ you can have

a Reuleaux polygon with n sides, and the fifty pence piece is indeed a Reuleaux heptagon. This is convenient for slot machines: it means that you can have a slot into which the coin only just fits, however you push it in.

What about the best shape for a conveyor belt? If we construct it in the obvious manner, then one of its two sides will be exposed and the other not. Eventually, the exposed side will wear out, but the other side will be in pristine condition, since it will not have been used at all. However, as any mathematician will tell you, not all surfaces have two sides. The most famous example of a one-sided surface is the MÖBIUS STRIP [IV.7 §2.3], which is obtained from a flat strip of paper by twisting one end through 180 degrees and joining it to the other end. If you have a long enough conveyor belt for it to be practical to give it a twist somewhere, then you can wear out both sides equally (this makes sense locally even if globally the belt now has just one side), thereby doubling the use you get out of the belt. (You might think it simpler just to turn the belt over after a while, but the Möbius-strip design has been taken seriously enough to be patented, and similar designs have been used as typewriter ribbons and in tape recorders.)

3 Scaling and Chirality

Why are Arctic mammals large? Is it just a fluke that they have evolved that way? This does not sound like a mathematical question, but some simple mathematics can easily convince us that it is not a fluke at all. Since the Arctic is cold and animals need heat, animals that are better at preserving heat are more likely to thrive. The rate at which an object loses heat is proportional to its surface area, but the rate at which it *generates* heat is proportional to its volume. So if you double the size of an animal in every direction, then the rate at which heat is generated goes up by a factor of eight, while the rate at which it is lost goes up by a factor of only four. That is, larger animals find it easier to preserve heat.

But why, in that case, are Arctic animals not much bigger still? This can be explained by a similar scaling argument. If you scale an animal up by a factor of t , then its volume, and hence its weight (animals, being made predominantly of water, tend to have roughly the same density), will multiply by t^3 . The animal has to support this weight with its bones. The amount of force you need to snap a bone is roughly proportional to the area of a cross-section of that bone, and areas go up by a factor of t^2 . So if t is too large, the animal will not be able to support its own weight. It does have the option

of increasing the relative thickness of its bones, but if t is very large then its legs will be too thick for this to be a practical solution.

A similar sort of scaling argument explains why, if you drop a mouse down a 1000 foot mine shaft, then, to quote Haldane, “on arriving at the bottom, [it] gets a slight shock and walks away.” In this case, air resistance is roughly proportional to surface area, while the gravitational pull is proportional to mass, and therefore to volume. It follows that, the smaller you are, the smaller your terminal velocity, and the less you are bothered by a fall.

A simple fact with many scientific ramifications is that two shapes can be reflections of each other without being rotations or translations. For example, if you see a hand without seeing the body to which it is attached, then you can tell whether it is a right hand or a left hand. (If you can shake it naturally with your right hand, then it is a right hand.) This phenomenon is known as *chirality*: a shape is *chiral* if it cannot be obtained from its mirror image by rotation or translation.

The notion of chirality appears in many parts of science. For example, many elementary particles have a fundamental property known as “spin,” which means that they often have right-handed versions and left-handed versions. In pharmacology, it is now understood that many molecules are chiral, and that the two different versions can have radically different properties. An example that had tragic consequences is the drug thalidomide: one form of it is effective against morning sickness while the other causes birth defects. Unfortunately, in the late 1950s several thousand pregnant women were given a 50:50 mixture of the two forms. Less harmful examples of the importance of chirality abound. For instance, there are many chemicals that smell or taste different when you look at their reflected versions. (This may seem paradoxical, but the explanation is simple: the sensors in our noses and mouths also contain molecules with chirality.)

So far we have been considering rigid motions, but some shapes are chiral in the stronger sense that not even a *continuous* motion in space is enough to turn them into their mirror images. Two interesting examples are the trefoil knot (see KNOT POLYNOMIALS [III.46]), which comes in a “right-handed” and a “left-handed” version (the proof that these two versions are genuinely distinct is not at all easy), and the Möbius strip, which was mentioned earlier. The rough reason that the Möbius strip is chiral is that when you do the twist, you do it either according to the “corkscrew

rule”—that is, twisting it as if you were pushing a corkscrew into the cork—or the opposite way. If you try to visualize it, you may be able to convince yourself that the direction of twist is not altered by continuous deformations, and also that the mirror image of a Möbius strip that obeys the corkscrew rule is a Möbius strip that does not obey the corkscrew rule.

4 Hearing Numerical Coincidences

Legend has it that Pythagoras, passing a blacksmith hammering a set of iron bars in a particularly pleasing way, was led to discover the laws of harmony. In modern terms, these laws say that two sounds go together particularly well (at least in the European tradition) if their frequencies are in the ratio r to s for some pair of small integers r and s : the smaller the better. As a result, people have tried to devise musical scales that have as many of these pleasing intervals as possible.

Unfortunately, there are limits to how well you can do. If you take a very simple ratio such as $3/2$, which corresponds to what musicians call a perfect fifth, then its powers— $9/4$, $27/8$, $81/16$, and so on—get successively more complicated. However, by great good fortune it happens that 2^{19} is rather close to 3^{12} . To be precise, $2^{19} = 524\,288$ and $3^{12} = 531\,441$, which is a difference of about 1.4%. It follows that $(3/2)^{12}$ is close to 2^7 . Since doubling a frequency raises the note by an octave, this says that twelve perfect fifths make an interval close to seven octaves. This allows one to build up a scale in which the fifths are *approximately* perfect.

There are many ways of doing the approximation. Early choices of musical scale would make some of the fifths perfect, at the expense of others. The modern compromise adopted by Western music for the last 250 years is to distribute the inaccuracies equally. If successive notes in a musical scale have frequencies in the ratio 1 to α , then starting from a frequency u the notes will have frequencies u , αu , $\alpha^2 u$, and so on. If you want k notes in the scale, then α^k should equal 2 (so that after k steps you have gone up by an octave). This means that all smaller powers of α must be irrational, so that all the other intervals in the scale are inharmonic! However, when $k = 12$, the fact that 3^{12} and 2^{19} are close has the consequence that α^7 , which equals $2^{7/12}$, is close to $3/2$ (more precisely, it is just over 1.4983), which means that all the fifths are close to perfect.

Tuning systems are discussed in more detail in MATHEMATICS AND MUSIC [VII.13 §2].

5 Information

Few things illustrate better how the abstract mathematical theory of one generation can become the common sense of the next than the following two closely related ideas: that all information can be expressed as a series of 0s and 1s, and that the “quantity of information” carried by a book, a picture, or a sound is proportional to the number of 0s and 1s required to express it.

A famous theorem of Shannon (described in RELIABLE TRANSMISSION OF INFORMATION [VII.6 §3]) tells us that the rate at which information can be transferred by signals depends on the range of frequencies available. For example, it is the change from signaling electrically along copper wires (with a narrow range of frequencies) to signaling by light (with a very wide range) that has allowed the massive data transfers required by the Internet. The sound waves we hear belong to a very narrow range of frequencies, while the light waves that we see belong to a wide range, and this is why we need much more memory on our computers to store an hour of film than an hour of music. Similarly, it may feel as though visual perception is a passive process—we point our eyes in a certain direction, they behave a bit like video cameras, and we just watch the video—but because light carries so much information, our brains actually have to resort to a wide variety of tricks to deal with it. What we think we see is actually a theatrical representation of reality that our brains have cunningly manipulated. This is why there are optical illusions, and why they continue to work even when you know how they work. By contrast, since sound carries so little information, our brains can process it in a much more direct way (though still not completely direct—there are aural illusions too, and the brain has tricks that help us to pick out the information we are actually interested in from all the sound waves that enter our ears).

When information is transmitted, there are almost always faults in the transmission system, so that our messages are not transmitted perfectly. How do we then recover the messages? Here is a Victorian parlor trick that shows how in a very simple case. One begins by writing down all sequences (x_1, x_2, \dots, x_7) such that every x_i is either 0 or 1 and such that the numbers $x_1 + x_3 + x_5 + x_7$, $x_2 + x_3 + x_6 + x_7$, and $x_4 + x_5 + x_6 + x_7$ are all even. An example of such a sequence is $(0, 0, 1, 1, 0, 0, 1)$.

If you think of these sequences as vectors in the vector space \mathbb{F}_2^7 (that is, the seven-dimensional space

where the scalars belong to the field of integers mod 2), then you will readily convince yourself that these three properties of a sequence are independent linear conditions, so the set of sequences in question is a four-dimensional subspace of \mathbb{F}_2^7 . Therefore, there are sixteen such sequences. A member of the audience is asked to take one of them and change it in one place. The magician can at once identify which digit has been changed. Let us see how this works if we change the third digit of the sequence above, so we now have the sequence $(y_1, \dots, y_7) = (0, 0, 0, 1, 0, 0, 1)$.

The first step is to note that $y_1 + y_3 + y_5 + y_7$ and $y_2 + y_3 + y_6 + y_7$ have become odd, while $y_4 + y_5 + y_6 + y_7$ is still even (since it is y_3 that has changed). Now the only number that belongs to the first two of the sets $\{1, 3, 5, 7\}$, $\{2, 3, 6, 7\}$, and $\{4, 5, 6, 7\}$ but not the third is 3. This tells us that x_3 is the variable that has been changed. How are the sets chosen so that this sort of argument will always work? The answer becomes clearer if we use the binary representations of the integers instead and put in a couple of leading zeros. Then the sets are $\{001, 011, 101, 111\}$, $\{010, 011, 110, 111\}$, and $\{100, 101, 110, 111\}$ and we see that the i th set is the set of integers with a 1 in the i th digit from the end. So if we know which of the three parities have been changed, then we know the binary representation of the place where the sequence was altered. Therefore, we can reconstruct the original sequence.

This trick, rediscovered by Hamming, is the ancestor of all the error-correcting methods (also discussed in RELIABLE TRANSMISSION OF INFORMATION [VII.6]) that allow our CDs and DVDs to perform flawlessly even if they are slightly scuffed.

The fact that there is a precise mathematical way of measuring information content is of considerable importance in genetics. It has been suggested that the amount of information carried by our DNA, though very large, is much smaller than the information required to describe our bodies completely. This would explain what experimental evidence also corroborates: that the DNA carries a set of general instructions, but the fine detail of our anatomy, such as our fingerprints and the precise arrangements of our capillaries, is partly a matter of chance. So, for example, if it were possible to rerun the growth of the fertilized egg that ended up as you, the result would be broadly similar to you, but small environmental differences would result in a different set of fingerprints and a different arrangement of capillaries.

Under certain circumstances, it is not enough just to transmit information: it must also be protected. If we send our credit card number over the Internet, we want to do so in such a way that it would be very hard for an eavesdropper to find that number. A mathematical way of doing this is described in CRYPTOGRAPHY [VII.7 §5].

Here is a slightly different but closely related problem. Suppose that Albert has a secret that he would like to share with Bertha (and only Bertha) in a conversation that everyone can hear. What is he to do? A first step is to think of *any* piece of information that they can share secretly—it turns out to be a short step from this to sharing a particular piece of information. The following procedure achieves this. First, Albert shouts out a large integer n and an integer u . Next, he chooses a large integer a , which he keeps secret (including from Bertha—obviously, since he does not yet know how to share secrets with her), and shouts out the value of u^a modulo n . Bertha then chooses an integer b , which she keeps secret, and shouts out the value of u^b modulo n . Now Albert is in a position to work out $u^{ab} = (u^b)^a$ modulo n , since Bertha has told him u^b and he knows a . Similarly, Bertha can use her secret number to work out $u^{ab} = (u^a)^b$ modulo n . Albert and Bertha now both know the number u^{ab} modulo n . This is a good example of a shared secret, because all that the eavesdroppers know is u^a , u^b , and n , and when n is large there is no known way of calculating u^{ab} modulo n from u^a and u^b modulo n , apart from methods that take far too long to be practical.

Now suppose that Albert wants to send a credit card number N to Bertha. Assuming that $1 \leq N \leq n$, then all he needs to do is shout out the number $u^{ab} + N$ modulo n . Bertha then subtracts the secret number u^{ab} and obtains N . (Albert should convey only one secret this way, or he will reveal information. For instance, if he sent another credit card number M using the same u^{ab} , then the eavesdroppers would know the value of $M - N$. But if he and Bertha choose new numbers n , u , a , and b and use those to share the value of M , the eavesdroppers will effectively know nothing about the pair (M, N) .)

Why do we believe that it is “hard” to calculate u^{ab} from u^a and u^b ? What if tomorrow somebody discovers a simple trick for doing it? Surprisingly, even though we cannot be absolutely sure that the problem is hard, there are very precise ways of discussing the question. In particular, there are extremely plausible conjectures, the truth of which would imply that it really is impossible to calculate u^{ab} in a short time.

These issues are discussed in considerable detail in COMPUTATIONAL COMPLEXITY [IV.20].

6 Mathematics in Society

A street in which all houses have front gardens is much prettier than a street in which all those front gardens have been converted into parking places. For some people, aesthetics are more important than convenience, so the effect of converting all the front gardens in a street may well be to reduce the values of all the houses. However, if you convert just one front garden, then it will increase the convenience for that household without making too much of a difference to the look of the street, so the value of that house will increase and the values of all the other houses will decrease slightly. Thus, for each individual house owner there is a financial incentive to convert the front garden, even though if everybody does so then everybody will lose financially.

Clearly, to avoid this unfortunate result the households must cooperate. Nash has shown how, starting from simple assumptions about fairness, there must be a system of mutual payments—for example, a household that wishes to convert its front garden might have to pay a charge that was shared between the other households—which will change their incentives in such a way that they will no longer want to ruin the street.

If the households do not wish to cooperate, Nash has shown that they come to a (usually less favorable) agreement which it is not in the interest of any single individual to break. A simple example of a situation in which no single individual may wish to change but a group acting in concert may wish to change is given by the following game. Suppose that three people hand to an umpire an envelope containing either the word “yes” or the word “no.” If two players have written the same thing and the third has not, then those two players get \$400 each and the third player gets nothing. However, if all three have written the same, then all three players get \$300. Suppose that the players meet before the game and agree that they will all write “yes” (in order to maximize their average gain). Then no single player will gain by writing “no” instead, but if two players decide to change then they will both gain.

Nash’s ingenious argument starts with an agreement that is not necessarily in equilibrium, and allows the parties to the agreement to modify their actions very slightly in a way that would improve their own situation if nobody else changed *their* actions. (However,

since the other parties *are* changing their actions, the total change may be preferable to nobody.) This results in a function that takes agreements to agreements. This function turns out to obey the conditions of THE KAKUTANI FIXED-POINT THEOREM [V.13 §2], from which it follows that there is an agreement that no single individual wishes to change. (See MATHEMATICS AND ECONOMIC REASONING [VII.8], particularly section 4, for a further discussion of Nash’s theorem. Another situation where individual and collective self-interest do not necessarily coincide is the flow of traffic (see THE MATHEMATICS OF TRAFFIC IN NETWORKS [VII.4 §4]).)

Not all applications of mathematical thought to social problems have such satisfactory outcomes. Suppose that there is to be an election (or, more generally, that society has to make a choice between various possibilities) with n candidates and m voters. Let us use the term “voting system” to mean any method of putting the n candidates in order given the preferences of the individual voters. Kenneth Arrow has shown that, under normal circumstances, there is no good voting system. More precisely, he has identified a small set of very reasonable sounding properties that one would wish a voting system to have, and shown that no voting system has all these properties. To give two examples of these properties, it is surely desirable that the final ranking of the candidates should depend on more than just the ranking of one individual voter, and one would also expect that if every voter prefers one candidate x to another candidate y , then x should be ranked higher than y . Instead of listing the other properties, we present a simpler result, known as Condorcet’s paradox, that gives some of the flavor of Arrow’s theorem. (Indeed, Arrow’s theorem can be regarded as a descendant of Condorcet’s paradox.) Consider three voters A, B, and C with the following preferences.

	A	B	C
First preferences	x	y	z
Second preferences	y	z	x
Third preferences	z	x	y

Observe that the majority of the voters prefer x to y , a majority prefer y to z , and a majority prefer z to x . Therefore, majority preference is not a TRANSITIVE RELATION [I.2 §2.3]. One consequence of this is that if voters are first asked to vote between two of x , y , and z and there is then a run-off between the winner of the

first vote and whichever of x , y , and z is left, then the remaining candidate will always win.

Probability is another branch of mathematics that plays a central role in modern society. In earlier societies people worked until they died. Today people can stop working and live off their savings. You can, of course, just live off the interest of your savings but this means that you will die with a large sum unspent. Alternatively, you can assume that you will live a certain number of years and run down your savings, reaching zero at precisely the moment you expect to expire. This will not be satisfactory if you live longer than you expect. The solution is to make a bet with a wealthy corporation. You pay them your capital and in return they pay you a certain sum every year until you die. If you die early then they have won their bet, and if you die late then they have lost. By taking a large number of such bets and relying on results like the STRONG LAW OF LARGE NUMBERS [III.73 §4], the corporation can be almost certain of making a profit in the long run. In effect you have paid a certain amount to transfer the risk (from the financial point of view) that you might live a long time from yourself to the corporation.

One of the earliest ways for mathematicians to make money was to become actuaries—that is, advisers on the appropriate price for transfer of risk in the situation described above. Nowadays, all sorts of risk (Will next year's coffee crop fail? Will the euro fall against the dollar?) are bought and sold and have to be priced. A discussion of risk pricing in general can be found in THE MATHEMATICS OF MONEY [VII.9].

7 Conclusion

In the past, mathematics has had a dramatic impact on physics and engineering. At one time this led to hopes that biological and sociological phenomena would eventually come to be explained mathematically as well. Later, such hopes came to seem unrealistic: it was understood that these areas contain “emergent phenomena” that are not easily amenable to a reductionist approach and are therefore genuinely harder to describe mathematically than the phenomena studied in the “harder sciences.” However, mathematicians are now beginning to grapple with such phenomena: as even the simple examples in this article have shown, one can apply mathematics to many areas outside its traditional domain, and doing so can be extremely illuminating.

VIII.4 Numeracy

Eleanor Robson

1 Introduction

Most of this *Companion* is rightly concerned with the theories and practices of professional mathematicians. But all human beings have ideas about numbers, space, and shape, and ways of putting these ideas to use. It could be said that numeracy is to mathematics what literacy is to literature: everyday, routine application versus expert, elite innovation. But while literacy is now a wildly fashionable subject of academic study, the word “numeracy” is not even recognized by my mass-market word processor. Yet an array of interesting work has been done on nonprofessional mathematical concepts, practices, and attitudes. They range from historical studies and ethnographies to cognitive analyses and developmental psychologies, and cover such diverse periods and places as ancient Iraq, the pre-Columbian Andes, and the European Middle Ages, as well as many parts of the contemporary world. By surveying selected studies on five broadly construed topics in numeracy and artisanal mathematics, I hope to make the case in this essay that numeracy is as valuable a topic of academic research as professional mathematics on the one hand and literacy on the other.

Mathematics has rarely been considered part of the sociology or anthropology of knowledge, as it has often been assumed to stand outside culture. That is to say, many people have held the view that one can only *think mathematics*, not think *about* it. Furthermore, such work as has been done on the place of mathematics in culture is fragmented: mathematical thinking in the developed world has tended to be studied by sociologists, but in the developing world by anthropologists; historians of mathematics have mostly taken as their subject the literate mathematics of the professional elite, while psychologists have generally focused on the acquisition of numeracy, by adults and children.

But, as we shall see, the way that societies and individuals regard mathematics is strongly contingent on many environmental factors. Educational, linguistic, visual, and intellectual cultures all shape mathematical thinking in different ways. That is not to say that there are no constraints, however. Humans all share basic anatomical similarities that influence our ways of thinking: we are approximately symmetrical about one vertical axis, for instance, which gives rise to arguably innate

concepts of left and right, front and back. And we all have fingers and opposable thumbs and the ability to subitize (that is, to recognize the size of a small set without counting its individual members). This, Reviel Netz has argued, makes human beings uniquely good at manipulating small groups of small objects, which has given rise to sophisticated systems of accounting and coinage. We shall return to Netz's work later.

The examples in this essay have been selected from studies of three very different clusters of world cultures. The ancient Middle East and Mediterranean (Egypt and Mesopotamia, classical Greece and Rome) have strongly influenced modern global culture in a variety of ways. Most obviously, the Euclidean tradition has been central to Western educational ideals for centuries, along with the teaching of Latin. And while the languages and writings of ancient Egypt and Mesopotamia are essentially nineteenth-century rediscoveries, their cultural influence runs in deep undercurrents throughout Western thinking, having percolated through classical and biblical learning. We should not be surprised, then, to discover the familiar as well as the alien in the world's oldest evidence for numeracy and artisanal mathematics. By contrast, the cultures of the pre-Columbian Americas are important for their very lack of contact with the premodern old world and thus their isolation from modernity. Virtually extinguished by the European conquests of the sixteenth and seventeenth centuries, and yet structurally similar to many old-world societies, they give a useful sense both of the constraints on numerate practice and thinking and of their diversity. Finally, this article also draws material from studies of the contemporary Americas, both South and North, in an attempt to break down the traditional disciplinary boundaries between past and present and between the developed world and the developing world. Numeracy is a feature of all human culture, wherever and whenever we have lived, and this should be reflected in how it is investigated.

2 Number Words and Social Values

Number words are usually studied for their mathematical content. French, for instance, shows traces of a vigesimal system in words such as *quatre-vingt*, meaning "four twenties," while the English word *eighty* is clearly derived from "eight tens." But in all languages number words also have social values attached, especially the counting numbers and words for sets. This is a rather different phenomenon from mystical numerology such as that of Late Antique Neo-Pythagoreanism.

For instance, Nichomachus's book *The Theology of Arithmetic* (written in the second century B.C.E. but now known only from later summaries) assigned esoteric meanings to the first ten integers, understanding those numbers to represent fundamental attributes of the cosmos. But the social values of number words are often much more prosaic than that. English, for instance, has a variety of words for "group of three," each of which is applicable to a particular range of objects and has particular social connotations. "Threesome" is not a synonym of "trinity" in everyday language, just as in musical terminology "trio" does not have the same referents as "triad" or "triplet." There is nothing mystical or esoteric in the use of these words; it is simply that, in addition to their semantic content, these words also carry implicit qualitative information about the sort of objects that are being grouped (sexually active adults, divine beings, musicians, musical notes, criminals, babies), about which society and individuals tend to form value judgments.

That numbers have a "social life" was first recognized by Gary Urton in his ethnographic study of the Quechua-speaking inhabitants of the Bolivian Andes. Structurally, Quechua numeration is straightforwardly decimal, much like modern European number systems, and is written with Arabic numerals. This has ensured its survival side by side with Spanish, but the fact that it is not particularly exotic relative to Western norms has caused it to be somewhat neglected academically. However, as Urton shows, there are two predominant social aspects to Quechua numeration: family relations on the one hand, and the idea of completeness or "rectification" on the other. There are also clear boundaries around what may be counted and who may count them.

All Quechua number words are composed of a dozen basic lexemes—one to ten, hundred, and thousand—which may be combined additively or multiplicatively, just as in English the word *thirteen* means "three and ten" and *thirty* means "three tens." Also as in English, Quechua number words tend to be a distinct lexical set; for instance, *kinsa* means "three" and nothing else. But where synonyms for cardinal numbers are fairly rare in English (one example is *dozen* for "twelve"), they are a normal part of Quechua speech:

- *iskaypaq chaupin*, "the middle of (sets of) twos," used of the third item in a group of five;
- *iskay aysana*, "double puller" (because the symbol 3 looks like two handles);

- *uquti*, “anus” (because the symbol 3 also looks like a human bottom);
- *uj yunta ch'ullayuaq*, “one pair, possessor of one standing alone” ($2 + 1 = 3$).

Family relations are most clearly visible in ordinal sequences, especially the names of the fingers, which are themselves important everyday counting tools. Urton lists six very similar sets of names, attested over the past 500 years. The most recent, collected by the Bolivian anthropologist Primitivo Nina Llanos in 1994, goes as follows:

- thumb, *mama riru*, “mother finger”;
- index finger, *juch'uy riru*, “small[er] finger”;
- middle finger, *chawpi riru*, “middle finger”;
- ring finger, *sullk'a riru*, “younger finger”;
- little finger, *sullk'aq sullk'an riru*, “younger sibling of the younger finger.”

Thus the thumb is considered both the oldest and the antecedent of the others and the little finger the youngest; this is true of all six attested variants of the finger names. The hands themselves are considered as two symmetrical halves of a unified whole—as are paired items in general. In Quechua, one hand alone (or indeed an odd number) is not in its natural state. As Urton explains:

[T]he motivation for *two* is the “loneliness” (*ch'ulla*) of *one*. “One” is an incomplete, alienated entity: it needs a “partner” (*ch'ullantin*). The principle and motivational force obtain ... regardless of whether the unit that composes the “one” is indivisible (e.g., a single digit) or divisible (e.g., a hand with five digits).

And more generally, Urton shows that in Quechua, odd numbers (*ch'ulla*) are incomplete while even numbers (*ch'ullantin*, “the part together with its pair”) represent the normal state of being.

But in Quechua society not everything is permissibly countable, even when there is no obvious difficulty in doing so. For example, they inventorize their herds, on whom they are often heavily economically dependent, not by counting but by naming. It is thought that counting individualizes the constituent members of the inseparable group, and thereby threatens its unity and fertility. If a herd *must* be counted then only a woman may do so; it is an unacceptably effeminate action for a man to carry out.

While restrictions on counting are not a notable feature of contemporary English-speaking culture, taboos

on particular numbers are still common. Why is thirteen considered so unlucky, for instance, particularly in North American hotels or on Fridays, while seven is regarded as lucky? In ancient Babylonia (modern-day southern Iraq) in the second and first millennia B.C.E., seven was thought to be particularly uncanny and unworldly. There were seven heavenly bodies (the Sun, the Moon, five visible planets), seven books of the *Epic of Creation*, and seven nights in each phase of the Moon. Demons, both beneficent and malevolent, were said to operate in groups of seven.

The Babylonians' primary numerical base for counting and recording groups of discrete objects was 60, factored into six groups of ten. The number 7 is, of course, the smallest one that is coprime to 60 and thus became a favorite subject of mathematics problems designed to be solved by trainee scribes. Further sexagesimal coprimes—11, 13, 17, 19—also featured prominently in ancient Babylonian mathematical problems and riddles. More often than not, however, the parameters were chosen in such a way that the tricky coprimes factored out or were otherwise disposed of, leaving an arithmetically innocuous answer:

I found a stone; I did not weigh it. I added a seventh. I added an eleventh. I weighed it: 1 mina. What was the original (weight of the) stone? The original stone was $\frac{2}{3}$ mina 8 shekels, $22\frac{1}{2}$ grains. (180 grains = 1 shekel; 60 shekel = 1 mina, ca. 0.5 kg.)

It is probably otiose to speculate whether the difficult mathematical properties of seven led directly to its cosmological demonization; the link is never made explicitly in any surviving cuneiform sources. But just as Babylonian demons failed to adhere to the norms of human behavior, so certain integers did not conform to the numerical patterns of the sexagesimally regular majority and the conceptual tools were not yet in place to explain that phenomenon in mathematical terms.

3 Counting and Calculating

While anyone can have views on whether particular numbers are lucky or unlucky, lonely or partnered, the ability to manipulate numbers arithmetically, and to take pleasure in doing so, is not universally shared. Both personal cognitive skills and social constraints are at work here. Patricia Cline Cohen argues that there were two key factors in the rapid rise in numerical competence in the early nineteenth-century United States. It was not that people suddenly became smarter. On the

one hand, the decimalization of money in the late eighteenth century meant that at last accountants, shopkeepers, and business owners were working with a single number base. At the same time, a new educational movement forsook the rote learning of arithmetical rules, applied mechanically to particular situations, for inductive instruction that encouraged pupils to calculate with fingers and counters, and in their heads, before they progressed to pen and paper. In this way some basic structural impediments were removed, both to the learning of number relationships, and to their application in commercial life.

Because modern decimal notation is a calculating system as well as a recording device, it is easily forgotten that other methods are just as effective. Indeed, for most communities, most of the time, numerals were simply a means to record the outcome of operations performed on the body or with other calculating tools. Finger counting and abacus use remained ubiquitous in the medieval Islamic world and Christian Europe long after knowledge of decimal numerals, together with AL-KHWARIZMĪ's [VI.5] treatise on how to use them, and cheap paper on which to write them, began to spread outward from Baghdad in the ninth century C.E. Their retention was not a knee-jerk reaction in the face of an overwhelmingly superior technology; rather, it took into account such factors as portability, speed of use, and a long-established trust in and institutional sanction of the old methods.

Indeed, it is difficult to overestimate just how old abacus calculation is. Reviel Netz identifies two evolutionary prerequisites for what he calls "counter culture," by which he means the uniquely and ubiquitously human use of small objects to represent other objects that are being counted, in one-one or one-many relationships. One is physiological: one needs to be able to pick up and manipulate small objects such as pebbles or shells. All primates share this ability thanks to prehensile fingers and opposable thumbs. The other is cognitive: one must be able to subitize, or recognize the size of a small set of up to about seven objects, without counting them individually. Stringed-bead abacuses exploit this most obviously, whether in the Russian-style ten-bead variety, whose fifth and sixth beads are always a different color from the others, or in the Japanese version, whose strings contain just four unit-beads and one five-bead each.

But, as Netz so powerfully puts it, "The abacus is not an artefact: it is a state of mind." All one needs is a

flat surface and a pile of small objects to act as counters. This extreme ephemerality makes the use of abacuses almost impossible to detect in the archaeological record, except in the rare cases where abacus counters can be recognized as such. Denise Schmandt-Besserat has argued that a sophisticated accounting system was developed in the Neolithic Middle East from the ninth millennium B.C.E. She proposes that the tiny, unbaked pieces of clay, crudely shaped into various simple geometrical figures and found in preliterate archaeological contexts from eastern Turkey to Iran, are ancient accounting tokens. It is certainly true that the earliest written numerals in the area, from southern Iraq in the late fourth millennium B.C.E., are marks on clay tablets that look remarkably like stylized impressions of such objects, and are visually distinct from the signs for the objects that were being counted, which were scratched onto the clay rather than impressed. It is also true that these earliest written records are almost exclusively accounting records, drawn up by temple administrators in the management of assets such as land, labor, and agricultural products. And from the fifth millennium B.C.E. onward, those tiny clay tokens are found in archaeological contexts—sealed into jars, for instance, or wrapped in little clay bundles, or carefully piled in the corners of storerooms—that are entirely compatible with their use as abacus counters. But Schmandt-Besserat's claim for a universally standardized system across the Middle East from several millennia before then is not provable: there is no way of establishing that they were not sometimes gaming pieces, for instance, or sling shot, or any number of other possibilities, and certainly no way of determining what specific shapes signified and to whom.

In fact, ad hoc means of counting and measuring are still everyday occurrences in all our lives, even among those with a high level of formal mathematics education. A team of anthropologists and psychologists, headed by Jean Lave, observed newcomers to a Californian Weight Watchers scheme in the 1980s as they adjusted to careful quantification of the food they were allowed to consume on the diet. One participant, who had taken a calculus course at college, was asked to modify a recipe calling for two-thirds of a cup of cottage cheese so that it contained three-quarters of that amount. Lave recalls: "He filled a measuring cup two-thirds full of cottage cheese, dumped it out on a cutting board, patted it into a circle, marked a cross on it, scooped away one quadrant, and served the rest." She comments:

Thus, “take three-quarters of two-thirds of a cup of cottage cheese” was not just the problem statement but also the solution to the problem and the procedure for solving it. The setting was part of the calculating process and the solution was simply the problem statement enacted within the setting. At no time did the Weight Watcher check his procedure against a paper and pencil algorithm, which would have produced $\frac{3}{4} \times \frac{2}{3} = \frac{1}{2}$ cup. Instead, the coincidence of problem, setting, and enactment was the means by which checking took place.

In other words, there are many situations in many people’s lives in which potentially applicable literate, school-taught mathematical procedures are ignored in favor of equally effective nonliterate ones that produce the correct result with the tools at hand. Numeracy takes many forms, not all of which entail writing.

4 Measurement and Control

The Weight Watcher invented a system of cottage-cheese measurement that satisfied him in its accuracy and fulfilled his immediate culinary needs. But as individuals and social groups we also accept the accuracy and consistency of standardized measurement systems, and the institutional necessity of counting and measuring particular things but not others. Theodore Porter has written eloquently of the twentieth century’s growing “trust in numbers,” whether of census statistics or environmental data. But institutionally sanctioned quantification is often contested, and it frequently alters the very phenomenon that is being pinned down. Cohen’s description of nineteenth-century North America is more generally apposite:

What people chose to count and measure reveals not only what was important to them but what they wanted to understand and, often, what they wanted to control. Further, how people counted and measured reveals underlying assumptions about the subject under study, assumptions ranging from plain old bias... to ideas about the structure of society and of knowledge. In some cases, the activity of counting and measuring itself altered the way people thought about what they were quantifying: numeracy could be an agent of change.

Cohen and Porter both explore problems raised by early nineteenth-century census taking. Porter describes the obstacles that the under-resourced Bureau de Statistique faced in obtaining accurate population data in post-revolutionary France. Without resorting to the old class categorizations of the *ancien régime*, it

needed to acknowledge the huge diversity of occupations and social structures across the country. To do so it relied on local officials to return a mass of quantitative data that was simply not readily available—and so the prefectures commissioned qualitative descriptions of their regions instead. As Porter puts it, in 1800 “France was not yet capable of being reduced to statistics.” Cohen analyzes the U.S. Census of 1840, which appeared to demonstrate a much higher rate of insanity among the black population in the abolitionist northern states than in the south. Pro-slavery factions took this as irrefutable evidence that slavery suited the black population much better than freedom did; abolitionists queried the trustworthiness of the census itself. Whether or not one chose to believe the data was more or less a matter of what one’s preexisting political convictions were. As Cohen shows, the source of the error lay in clumsily designed recording sheets, in which the “idiot white” and “idiot black” columns were easily confused, resulting in the misrecording of many elderly senile inhabitants of all-white households. In the 1840s, however, the public debate was not about methodology, but whether fraud had been committed: the numbers themselves could not lie.

Two thousand years earlier, as Serafina Cuomo has shown, the Roman land surveyor Frontinus opined that the world was essentially unknowable without quantitative intervention, and that the trustworthiness of that measure was dependent on professional expertise:

*The basis of the art of measuring lies in the experience of the agent. It is in fact impossible to express the truth of the places or of the size without calculable lines, because the wavy and uneven edge of any piece of land is enclosed by a boundary which, because of the great quantity of unequal angles, can be contracted or expanded, even when their number [that is, the number of the angles] remains the same. Indeed pieces of land which are not finally demarcated have a fluctuating space and an uncertain determination of *iugera*.*

The natural world is problematically irregular, Frontinus believed, and must be disciplined into quantified straight lines—and, ideally, marked out into grids of 2400 foot squares (*iugera*)—in order to be brought under control. The Roman reshaping of the landscape through its quantification is still visible throughout Europe, the Middle East, and North Africa today, both on land and from the air.

The Incas, by contrast, brought time, space, society, and the gods under control through radial lines

in the landscape, tied to the ceremonial year. Before Spanish-led Christianization in the sixteenth-century, the heart of the Inca cosmos was the sacred city of Cuzco in the Peruvian Andes. The Incas divided the world into unequal quarters or *tawantinsuyu* “the four parts together,” radiating out from the Temple of the Sun. Through each *suyu* ran nine to fourteen *ceque* paths through the mountains, forty-one in total, with an average of eight *huaca* shrines stationed on each. The local inhabitants performed a ritual at one of the 328 *huacas* every day of the sacred year (composed of twelve months of $27\frac{1}{3}$ days). Thus the religious focus of the Inca state moved systematically around its territory, day by day and from community to community, binding every social group into the same calendar, cult, and cosmos.

Numeracy, then, is a powerful institutional tool: measuring, quantifying, and classifying can transform an unknowable mass of individual people, places, or things into manageable categories of known entities—and this institutionally imposed structure in turn shapes the self-identities of those being managed. Institutional numeracy, while imposed from above, is always dependent to some degree on community-wide support and cooperation, if not necessarily for the objects of account then always for the counters. Attempts at censuses in the eighteenth century did not fail because people refused to be reduced to numbers in boxes, but because those charged with collecting the data had neither the infrastructural means to do so nor an intellectual outlook that valued quantification. Inca and Roman societies, by contrast, were able to produce whole classes of the professionally numerate who did.

5 Numeracy and Gender

In modern anglophone culture, academic mathematics is popularly considered a male pursuit—and women supposedly have to subordinate or compromise their femininity if they are to succeed in it. But such perceptions are far from universal: studies collected by Barbro Grevholm and Gila Hanna, for instance, show that in the early 1990s some 80% of Kuwaiti and over half of Portuguese undergraduate mathematics majors were women. However, as the following examples demonstrate, this has more to do with how particular societies construct ideals of femininity and masculinity and with what they count as mathematical activity than with any intrinsically gendered properties of mathematics itself.

For most of the second millennium B.C.E., Babylonian scribes understood professional numeracy to be

a divine gift—not from the gods in general but from a handful of powerful goddesses. In the literary works that scribal students memorized as part of their professional training, creator gods bestowed land-measuring equipment and numeracy on those goddesses to enable them to manage household estates equitably. In a myth now known as *Enki and the World Order* the great god Enki announces:

My illustrious sister, holy Nisaba,
Is to receive the 1-rod measuring reed.
The lapis lazuli rope is to hang from her arm.
She is to proclaim all the great divine powers.
She is to fix boundaries and mark borders. She is
to be the scribe of the Land.
The gods' eating and drinking are to be in her
hands.

The scribes' literary works also portrayed Nisaba as the patron of institutional numeracy in the real world: she in turn provided mensuration tools to scribes and kings to enable them to uphold justice in society.

Another scholastic literary genre was the scribal dialogue, in which the protagonists argue over the ideals of scribal professionalism. In one such debate the young scribe Enki-manshum explicitly relates metrological competence to social justice:

When I go to divide a plot, I can divide it; when
I go to apportion a field, I can apportion the
pieces,
So that when wronged men have a quarrel I
soothe their hearts and ...
Brother will be at peace with brother, their
hearts ...

This was not merely a literary trope: law codes promulgated by real-life Babylonian kings often began with prologues claiming that they would uphold fairness in commercial measuring, weighing, and counting, and included provisions for punishing metrological fraud. Many hundreds of legal records survive, attesting to the settlement of land disputes through accurate professional measurement and calculation. In the nineteenth-century B.C.E. city of Sippar, the judges who held court in the temple of Shamash, god of justice, employed female scribes and surveyors as well as male (often from the same families). Further, the personal seals of fourteenth-century B.C.E. royal land surveyors were often dedicated to Nin-sumun, the divine mother of the legendary hero Gilgamesh: for them the numerate goddess who bestows numerate justice was no school story but at the very heart of their professional self-identity.

Terri: very tricky to know what to do with this sentence that the proofreader pointed out wasn't great before. Tim has suggested this version - OK with you?

In ancient Babylonia, then, numeracy and metrology gained institutional authority and power as much through association with divine femininity as with royal masculinity. Many modern societies, by contrast, defeminize numerate thought and activity by denying its mathematical status when it is carried out by women. Gary Urton's study of Quechua numeration started out as an ethnography of Bolivian weaving, which, he discovered, was based on highly intricate symmetrical patterns that the (female) weavers know by heart. They count off threads effortlessly, unerringly picking up where they have left off after interruptions to nurse babies, prepare food, or attend to other domestic matters. And yet the men of the area categorically told Urton that the weavers "can't count"—because when a woman sells her finished weavings at market she will invariably ask another woman of the group to check her takings to ensure that she has not been cheated.

Urton was taught to weave by Irene Flores Condori, a twelve-year-old girl. He recalls:

On one occasion, a stern old woman ... asked me point blank if, by weaving, I was trying to be like a woman. I answered by telling her that in some villages I know of, it is the men rather than the women who do the weaving.... The old woman gave us both a wry look and asked, if that was the case, then is it the women in those villages who have the penises!

Weaving was such a strongly gendered activity that this and other incidents led Urton to feel that "my behavior was being tolerated to the degree that it was only because, as an outsider, I was not subject to the same rules and expectations as local men." Weaving is exclusively women's work and therefore its intrinsically numerate character is socially invisible; women are more reluctant than men to trust strangers to handle money fairly and are therefore considered innumerate.

Mary Harris shows how a similarly powerful gender divide developed in Victorian Britain as primary education became available to an ever-widening section of the populace. Mathematics was regarded as the quintessentially male school subject, while needlework was the epitome of femininity. Yet:

Every garment knitted to fit a particular body depends on the principle of ratio. Every pinafore pattern copied from a blackboard requires visual interpretation of scaling and the ability to draw a smooth curve. All the fine stitching that the early Inspectors were unable to tell from machine stitching depended on the ability to

judge equal distances by eye and maintain them in a straight row.

In other words, wherever girls and women weave, knit, or sew they are unwittingly engaging numerate aptitudes and skills, often highly creatively, just as Molière's Monsieur Jourdain had been speaking prose all his life "without knowing anything about it."

6 Numeracy and Literacy, School and Supermarket

Perhaps one reason that women's work is not often thought to belong to the realm of professional numeracy is that numeracy is so often considered (when it is considered at all) as a subset of literacy. As Reviel Netz puts it,

With Arabic numerals, numbers appear as secondary to writing, benefiting from tools that were largely invented to record verbal systems and not numerical symbols. In broad historical perspective, this is the exception and not the rule. The rule is that, across cultures, and especially in early cultures, the record and manipulation of visual symbols precede and predominate over the record and manipulation of verbal symbols.

Netz is thinking here of counters and abacuses, but the Bolivian weavers remind us that numeracy does not have to entail symbolic manipulation at all. One may count threads, llamas, ideas, anything, and perform calculations without the intervention of external tools. The use of fingers and other body parts has cropped up repeatedly in the examples presented in this essay. Much of the weavers' mental work is so naturalized within the rhythms and movements of their bodies that they can no longer verbalize the mental or physical processes involved. (That is why Urton chose a young girl as his teacher, who was still learning the craft, rather than a fully competent adult woman.) Nonliterate numerate practices and ideas, especially in the developing world, are often labeled by academic observers as "ethnomathematics." But this raises difficult questions about the appropriate use of the "ethno" prefix and about the border between numeracy and mathematics. How do we distinguish numeracy from mathematics, and where does ethnomathematics fit in?

When Ubiratan D'Ambrosio coined the term "ethnomathematics" in the mid 1970s it was to describe the study of mathematics "in direct relation to [its] social, economic, and cultural backgrounds," a subject lying

“on the borderline between the history of mathematics and cultural anthropology.” However, for many, particularly within mathematics education, it has come to mean the study of culturally “other” mathematics, as if only the academically marginalized have ethnicity (just as, according to some lazy academic views, only women have gender). This semantic shrinkage is doubly damaging, for it implies that “ethnic” cultures are not fully numerate, while rendering the mainstream of academic mathematics, both past and present, invisible to sociological, anthropological, or ethnographic research. Nor does it distinguish between the intellectual creativity that is mathematics and the routine application of numeracy.

If “ethnomathematics” is an unhelpful term, there are useful alternatives. An influential Brazilian study of childhood numeracy, by Terezinha Nunes and colleagues, distinguishes formally learned “school mathematics” from “street mathematics” created informally by the same children. Jean Lave’s ethnography of adult numeracy in 1980s California likewise contrasts “school arithmetic” with “supermarket arithmetic.” The participants in her study often described themselves as arithmetically incompetent and “were unaware of the efficacy of their math practice in the supermarket, and some did not know, even that they used arithmetic practices there.” Yet often the supermarket setting required the solution of mathematical problems of much greater complexity than superficially similar scholastic “word problems”:

The shopper was standing in front of a produce display. As she spoke she put apples, one at a time, into a bag. She put the bag in the cart as she finished talking: “There’s only about three or four [apples] at home, and I have four kids, so you figure at least two apiece in the next three days. These are the kinds of things I have to resupply. I only have a certain amount of storage space in the refrigerator, so I can’t load it up totally.... Now that I’m home in the summertime, this is a good snack food. And I like an apple sometimes at lunchtime when I come home.”

While explicitly considering such variables as the number of apple-consumers in the household, their rate of consumption, fridge storage space, and perhaps implicitly the apples’ price and probably shelf life, the shopper selected nine apples to buy. She might also have compared the prices of different varieties of apple and/or considered whether loose or prepackaged apples were the better buy—all typical supermarket activities that Lave and her researchers observed

and correlated with the same subjects’ performance in written tests of arithmetically similar skills. They found “not a single significant correlation between frequency of calculation in a supermarket, and scores on math test, multiple choice test or number facts.... Success and frequency of calculation in supermarket and simulation experiment bear no statistical relationship with schooling, years since schooling was completed, or age.”

Rather depressingly for educators, perhaps, Lave’s work suggests that training in school mathematics has little or no impact on numerical competence in adult life. (Interestingly, this finding conflicts with Cohen’s historical argument discussed above, relating improvements in mathematics education to rising standards of numeracy in early nineteenth-century North America.) Rather, as she and Étienne Wenger argue, learning takes place most effectively when it is situated in the social and professional context to which it pertains, through interaction and collaboration with competent practitioners, rather than through abstract, decontextualized classroom learning. Learners become part of a “community of practice” that inculcates not only the necessary technical skills but also the beliefs, standards, and behaviors of the group. Through gains in competence, confidence, and social acceptance, the learner moves from the periphery toward the center of the practice community, in due course becoming accepted as a fully fledged expert. It is perhaps in this light, then, that we should understand the process of becoming professionally numerate. But if situated learning is so effective, the development of supra-utilitarian educational mathematics in the societies of the ancient Middle East and Mediterranean is a major historical conundrum that has hitherto gone unrecognized.

7 Conclusions

This essay began by suggesting that “numeracy is to mathematics what literacy is to literature.” But the case studies presented here show that numeracy has a far greater cognitive reach than that. Throughout time and across the world countless individuals and societies have managed perfectly well, and continue to thrive, without writing; none has yet been attested without counting, measuring, or pattern-making in some form or other. In this light a better formulation might be that “numeracy is to mathematics what *language* is to literature.” Indeed babies, toddlers, and young children learn many essential mathematical skills through engagement with their immediate environment well before

formal school learning begins. Just as some children grow into more articulate adults than others, with or without highly developed skills in reading and writing, so they may become more or less numerate in their everyday practices, independently of their competence in school mathematics.

There are many deep and important questions about the relationships between numeracy and mathematics, language and literacy that have hardly yet been formulated, let alone explored: this is perhaps one of the most open fields of enquiry in academia today. This essay has only scratched the surface of a fascinating and complex subject that has paradoxically been overlooked because of its very ubiquity and centrality to human existence. In the next few decades, a wide range of interdisciplinary approaches will almost certainly yield important and surprising discoveries about numeracy that today we can only guess at.

Further Reading

- Ascher, M. 2002. *Mathematics Elsewhere: An Exploration of Ideas Across Cultures*. Princeton, NJ: Princeton University Press.
- Bloor, D. 1976. *Knowledge and Social Imagery*. London: Routledge & Kegan Paul.
- Cohen, P. C. 1999. *A Calculating People: The Spread of Numeracy in Early America*, 2nd edn. New York and London: Routledge.
- Crump, T. 1990. *The Anthropology of Numbers*. Cambridge: Cambridge University Press.
- Cuomo, S. 2000. Divide and rule: Frontinus and Roman land-surveying. *Studies in History and Philosophy of Science* 31: 189–202.
- D'Ambrosio, U. 1985. Ethnomathematics and its place in the history and pedagogy of mathematics. *For the Learning of Mathematics* 5:41–48.
- Gerdes, P. 1998. *Women, Art and Geometry in Southern Africa*. Trenton, NJ: Africa World Press.
- Glimp, D., and M. R. Warren, eds. 2004. *The Arts of Calculation: Quantifying Thought in Early Modern Europe*. Basingstoke: Palgrave Macmillan.
- Grevholm, B., and G. Hanna. 1995. *Gender and Mathematics Education: An ICMI Study in Stiftsgården Åkersberg, Höör, Sweden, 1993*. Lund: Lund University Press.
- Harris, M. 1997. *Common Threads: Women, Mathematics, and Work*. Stoke on Trent: Trentham Books.
- Lave, J. 1988. *Cognition in Practice: Mind, Mathematics and Culture in Everyday Life*. Cambridge: Cambridge University Press.
- Lave, J., and E. Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.
- Netz, R. 2002. Counter culture: towards a history of Greek numeracy. *History of Science* 40:321–52.
- Nunes, T., A. Dias, and D. Carraher. 1993. *Street Mathematics and School Mathematics*. Cambridge: Cambridge University Press.
- Porter, T. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.
- Robson, E. 2008. *Mathematics in Ancient Iraq: A Social History*. Princeton, NJ: Princeton University Press.
- Schmandt-Besserat, D. 1992. *From Counting to Cuneiform*. Austin, TX: University of Texas Press.
- Urton, G. 1997. *The Social Life of Numbers: A Quechua Ontology of Numbers and Philosophy of Arithmetic*. Austin, TX: University of Texas Press.

VIII.5 Mathematics: An Experimental Science

Herbert S. Wilf

1 The Mathematician's Telescope

Albert Einstein once said, “You can confirm a theory with experiment, but no path leads from experiment to theory.” But that was before computers. In mathematical research now, there’s a very clear path of that kind. It begins with wondering what a particular situation looks like in detail; it continues with some computer experiments to show the structure of that situation for a selection of small values of the parameters of the problem; and then comes the human part: the mathematician gazes at the computer output, attempting to see and to codify some patterns. If this seems fruitful, then the final step requires the mathematician to prove that the apparent pattern is really there, and is not a shimmering mirage above the desert sands.

A computer is used by a pure mathematician in much the same way that a telescope is used by a theoretical astronomer. It shows us “what’s out there.” Neither the computer nor the telescope can provide a theoretical explanation for what it sees, but both of them extend the reach of the mind by providing numerous examples that might otherwise be hidden, and from which one has some chance of perceiving, and then demonstrating, the existence of patterns, or universal laws.

In this article I would like to show you some examples of this process at work. Naturally the focus will be on examples in which some degree of success has been realized, rather than on the much more numerous cases where no pattern could be perceived, at least

by my eyes. Since my work is mainly in combinatorics and discrete mathematics, the focus will also be on those areas of mathematics. It should not be inferred that experimental methods are not used in other areas; only that I don't know those applications well enough to write about them.

In one short article we cannot even begin to do justice to the richly varied, broad, and deep achievements of experimental mathematics. For further reading, see the journal *Experimental Mathematics* and the books by Borwein and Bailey (2003) and Borwein et al. (2004).

In the following sections we give first a brief description of some of the useful tools in the armament of experimental mathematics, and then some successful examples of the method, if it is a method. The examples have been chosen subject to fairly severe restrictions:

- (i) the use of computer exploration was vital to the success of the project; and
- (ii) the outcome of the effort was the discovery of a new theorem in pure mathematics.

I must apologize for including several examples from my own work, but those are the ones with which I am most familiar.

2 Some of the Tools in the Toolbox

2.1 Computer Algebra Systems

The mathematician who enjoys using computers will find an enormous number of programs and packages available, beginning with the two major computer algebra systems (CASs), Maple and Mathematica. These programs can provide so much assistance to a working mathematician that they must be regarded as essential pieces of one's professional armamentarium. They are extremely user-friendly and capable.

Typically one uses a CAS in interactive mode, meaning that you type in a one-line command and the program responds with its output, then you type in another line, etc. This modus operandi will suffice for many purposes, but for best results one should learn the programming languages that are embedded in these packages. With a little knowledge of programming, one can ask the computer to look at larger and larger cases until something nice happens, then take the result and use another package to learn something else, and so forth. Many are the times when I have written little programs in Mathematica or Maple and then gone away for the weekend leaving the computer running and searching for interesting phenomena.

2.2 Neil Sloane's Database of Integer Sequences

Aside from a CAS, another indispensable tool for experimentally inclined mathematicians, particularly for combinatorialists, is Neil Sloane's "On-Line Encyclopedia of Integer Sequences," which is on the Web at www.research.att.com/~njas. At present, this contains nearly 100 000 integer sequences and has full search capabilities. A great deal of information is given for each sequence.

Suppose that for each positive integer n you have an associated set of objects that you want to count. You might, for example, be trying to determine the number of sets of size n with some given property, or you might wish to know how many prime divisors n has (which is the same as counting the set of these prime divisors). Suppose further that you've found the answer for $n = 1, 2, 3, \dots, 10$, say, but you haven't been able to find any simple formula for the general answer.

Here's a concrete example. Suppose you're working on such a problem, and the answers that you get for $n = 1, 2, \dots, 10$ are 1, 1, 1, 1, 2, 3, 6, 11, 23, 47. The next step should be to look online to see if the human race has encountered your sequence before. You might find nothing at all, or you might find that the result that you'd been hoping for has long since been known, or you might find that your sequence is mysteriously the same as another sequence that arose in quite a different context. In the third case, an example of which is described below in section 3, something interesting will surely happen next. If you haven't tried this before, do look up the little example sequence above, and see what it represents.

2.3 Krattenthaler's Package "Rate"

A very helpful Mathematica package for guessing the form of hypergeometric sequences has been written by Christian Krattenthaler and is available from his Web site. The name of the package is Rate (rot'-eh), which is the German word for "guess."

To say what a hypergeometric sequence is let's first recall that a rational function of n is a quotient of two polynomials in n , like $(3n^2 + 1)/(n^3 + 4)$. A hypergeometric sequence $\{t_n\}_{n \geq 0}$ is one in which the ratio t_{n+1}/t_n is a rational function of the index n . For example, if $t_n = \binom{n}{7}$ then t_{n+1}/t_n works out to be $(n+1)/(n-6)$, which is a rational function of n , so $\{t_n\}_{n \geq 0}$ is a hypergeometric sequence. Other examples

are

$$n!, (7n+3)!, \binom{n}{7}t^n, \frac{(3n+4)!(2n-3)!}{4^n n!^4},$$

all of which are easily seen to be hypergeometric.

If you input the first several members of the unknown sequence, Rate will look for a hypergeometric sequence that takes those values. It will also look for a hyper-hypergeometric sequence (i.e., one in which the ratio of consecutive terms is hypergeometric), and a hyper-hyper-hypergeometric sequence, etc.

For example, the line

Rate[1, 1/4, 1/4, 9/16, 9/4, 225/16]

elicits the (somewhat inscrutable) output

$$\{4^{1-i0}(-1+i0)!^2\}.$$

Here $i0$ is the running index of Rate, so we would normally write that answer as, say,

$$\frac{(n-1)!^2}{4^{n-1}} \quad (n = 1, 2, 3, 4, 5, 6),$$

which fits the input sequence perfectly. Rate is a part of the *Superseeker* front end to the Integer Sequences database, discussed in section 2.2 above.

2.4 Identification of Numbers

Suppose that, in the course of your work, you encountered a number, let's call it β , which, as nearly as you could calculate it, was 1.218041583332573. It might be that β is related to other famous mathematical constants, like π , e , $\sqrt{2}$, and so forth, or it might not. But you'd like to know.

The general problem that is posed here is the following. We are given k numbers, $\alpha_1, \dots, \alpha_k$ (the *basis*), and a target number α . We want to find integers m, m_1, \dots, m_k such that the linear combination

$$m\alpha + m_1\alpha_1 + m_2\alpha_2 + \dots + m_k\alpha_k \quad (1)$$

is an extremely close numerical approximation to 0.

If we had a computer program that could find such integers, how would we use it to identify the mystery constant $\beta = 1.218041583332573$? We would take the α_i to be a list of the logarithms of various well-known universal constants and prime numbers, and we would take $\alpha = \log \beta$. For example, we might use

$$\{\log \pi, 1, \log 2, \log 3\} \quad (2)$$

as our basis. If we then find integers m, m_1, \dots, m_4 such that

$$m \log \beta + m_1 \log \pi + m_2 + m_3 \log 2 + m_4 \log 3 \quad (3)$$

is extremely close to 0, then we will have found that our mystery number β is extremely close to

$$\beta = \pi^{-m_1/m} e^{-m_2/m} 2^{-m_3/m} 3^{-m_4/m}. \quad (4)$$

At this point we will have a judgment to make. If the integers m_i seem rather large, then the presumed evaluation (4) is suspect. Indeed, for any target α and basis $\{\alpha_i\}$ we can always find huge integers $\{m_i\}$ such that the linear combination (1) is *exactly* 0, to the limits of machine precision. The real trick is to find a linear combination that is extraordinarily close to 0, while using only "small" integers m, m_i , and that is a matter of judgment. If the judgment is that the relation found is real, rather than spurious, then there remains the little job of proving that the suspected evaluation of α is correct, but that task is beyond our scope here. For a nice survey of this subject, see Bailey and Plouffe (1997).

There are two major tools that can be used to discover linear dependencies such as (1) among the members of a set of real numbers. They are the algorithms PSLQ, of Ferguson and Forcade (1979), and LLL, of Lenstra et al. (1982), which uses their lattice basis reduction algorithm. For the working mathematician, the good news is that these tools are available in CASS. For example, Maple has a package, *IntegerRelations* [*LinearDependency*], which places the PSLQ and the LLL algorithms at the immediate disposal of the user. Similarly there are Mathematica packages on the Web that can be freely downloaded and which perform the same functions.

An application of these methods will be given in section 7. For a quick illustration, though, let us try to recognize the mystery number $\beta = 1.218041583332573$. We use as a basis the list in (2) above, and we put this list, augmented by $\log 1.218041583332573$, into the *IntegerRelations* [*LinearDependency*] package. The output is the integer vector $[2, -6, 0, 3, 4]$, which tells us that $\beta = \pi^3 \sqrt{2}/36$, to the number of decimal places carried.

2.5 Solving Partial Differential Equations

I had occasion recently to need the solution to a certain partial differential equation (PDE) that arose in connection with a research problem that was posed by Graham et al. (1989). It was a first-order linear PDE, so in principle the METHOD OF CHARACTERISTICS [III.51 §2.1] gives the solution. As those who have tried that method know, it can be fraught with technical difficulties relating to the solution of the associated ordinary differential equations.

Table 1 The first ninety-five values of $b(n)$.

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1	2	1	3	2	3	1	4	3	5	2	5	3	4	1	5	4	7
19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
3	8	5	7	2	7	5	8	3	7	4	5	1	6	5	9	4	11	7
38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56
10	3	11	8	13	5	12	7	9	2	9	7	12	5	13	8	11	3	10
57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
7	11	4	9	5	6	1	7	6	11	5	14	9	13	4	15	11	18	7
76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
17	10	13	3	14	11	19	8	21	13	18	5	17	12	19	7	16	9	11

However, some extremely intelligent packages are available for solving PDEs. I used the Maple command `pdsolve` to handle the equation

$$(1 - \alpha x - \alpha' y) \frac{\partial u(x, y)}{\partial x} = y(\beta + \beta' y) \frac{\partial u(x, y)}{\partial y} + (y + (\beta' + y')y)u(x, y)$$

with $u(0, y) = 1$. `pdsolve` found that

$$u(x, y) = \frac{(1 - \alpha x)^{-y/\alpha}}{(1 + (\beta'/\beta)y(1 - (1 - \alpha x)^{-\beta/\alpha}))^{1+y'/\beta'}}$$

is the solution, and that enabled me to find explicit formulas for certain combinatorial quantities, with much less work and fewer errors than would otherwise have been possible.

3 Thinking Rationally

The following problem appeared in the September/October 1997 issue of *Quantum* (and was chosen by Stan Wagon for the Problem of the Week archive).

How many ways can 90 316 be written as

$$a + 2b + 4c + 8d + 16e + 32f + \cdots,$$

where the coefficients can be any of 0, 1, or 2?

In standard combinatorial terminology, the question asks for the number of *partitions* of the integer 90 316

into powers of 2, where the multiplicity of each part is at most 2.

Let's define $b(n)$ to be the number of partitions of n , subject to the same restrictions. Thus $b(5) = 2$ and the two relevant partitions are $5 = 4 + 1$ and $5 = 2 + 2 + 1$. Then it is easy to see that $b(n)$ satisfies the recurrences $b(2n+1) = b(n)$ and $b(2n+2) = b(n) + b(n+1)$, for $n = 0, 1, 2, \dots$, with $b(0) = 1$.

It is now easy to calculate particular values of $b(n)$. This can be done directly from the recurrence, which is quite fast for computational purposes. Alternatively, it can be shown quite easily that our sequence $\{b(n)\}_0^\infty$ has the generating function

$$\sum_{n=0}^{\infty} b(n)x^n = \prod_{j=0}^{\infty} (1 + x^{2^j} + x^{2 \cdot 2^j}).$$

(For more information on generating functions, see ALGEBRAIC AND ENUMERATIVE COMBINATORICS [IV.18 §§2.4, 3], or see Wilf (1994).) This helps us to avoid much programming when working with the sequence, because we can use the built-in series-expansion instructions in Mathematica or Maple to show us a large number of terms in this series quite rapidly. Returning to the original question from *Quantum*, it is a simple matter to compute $b(90\,316) = 843$ from the recurrence. But let's try to learn more about the sequence $\{b(n)\}$ in general. To do that we open up our telescope, and calculate the first ninety-five mem-

bers of the sequence, i.e., $\{b(n)\}_0^{94}$, which are shown in table 1. The question is now, as it always is in the mathematics laboratory, what patterns do you see in these numbers?

Just as an example, one might notice that when n is 1 less than a power of 2, it seems that $b(n) = 1$. The reader who is fond of such puzzles is invited to cease reading here for the moment (without peeking at the next paragraph), and look at table 1 to spend some time finding whatever interesting patterns seem to be there. Computations up to $n = 94$ aren't as helpful for a quest like this as computations up to $n = 1000$ or so might be, so the reader is also invited to compute a much longer table of values of $b(n)$, using the above recurrence formulas, and to study it carefully for fruitful patterns.

OK, did you notice that if $n = 2^a$ then $b(n)$ appears to be $a + 1$? How about this one: in the block of values of n between 2^a and $2^{a+1} - 1$, inclusive, the largest value of $b(n)$ that seems to occur is the Fibonacci number F_{a+2} . There are many intriguing things going on in this sequence, but the one that was of crucial importance in understanding it was the observation that *consecutive values of $b(n)$ seem always to be relatively prime*.¹

It was totally unexpected to find a property of the values of this sequence that involved the multiplicative structure of the positive integers, rather than their additive structure, which would have been quite natural. This is because the theory of partitions of integers belongs to the additive theory of numbers, and multiplicative properties of partitions are rare and always cherished.

Once this relative primality is noticed, the proof is easy. If m is the smallest n for which $b(n)$, $b(n + 1)$ fail to be relatively prime, then suppose $p > 1$ divides both of them. If $m = 2k + 1$ is odd, then the recurrence implies that p divides $b(k)$ and $b(k + 1)$, contradicting the minimality, whereas if $m = 2k$ is even, the recurrence again gives that result, finishing the proof.

Why was it so interesting that consecutive values appeared to be relatively prime? Well, at once that raised the question of whether every possible relatively prime pair (r, s) of positive integers occurs as a pair of consecutive values of this sequence, and if so, whether every such pair occurs once and only once. Both of those possibilities are supported by the table of values above, and upon further investigation both turned out to be true. See Calkin and Wilf (2000) for details.

1. Two positive integers are *relatively prime* if they have no common factor.

Figure 1 The Ferrers board.

The bottom line here is that *every positive rational number occurs once and only once, and in reduced form, among the members of the sequence $\{b(n)/b(n + 1)\}_0^\infty$* . Hence the partition function $b(n)$ induces an enumeration of the rational numbers, a result which was found by gazing at a computer screen and looking for patterns.

Moral: be sure to spend many hours each day gazing at your computer screen and looking for patterns.

4 An Unexpected Factorization

One of the great strengths of computer algebra systems is that they are very good at factoring. They can factor very large integers and very complicated expressions. Whenever you run into some large expression as the answer to a problem that interests you, it is good practice to ask your CAS to factor it for you. Sometimes the results will surprise you. This is one such story.

The theory of *Young tableaux* forms an important part of modern combinatorics. To create a Young tableau we choose a positive integer n and a partition $n = a_1 + a_2 + \cdots + a_k$ of that integer. We'll use the integer $n = 6$ and the partition $6 = 3 + 2 + 1$ as an example. Next we draw the *Ferrers board* of the partition, which is a truncated chessboard that has a_1 squares in its first row, a_2 in its second row, etc., the rows being left-justified. In our example, the Ferrers board is as shown in figure 1.

To make a tableau, we insert the labels $1, 2, \dots, n$ into the n cells of the board in such a way that the labels increase from left to right across each row and increase from top to bottom down every column. With our example, one way to do this is as shown in figure 2.

One of the important properties of tableaux is that there is a one-to-one correspondence, known as the Robinson-Schensted-Knuth (RSK) correspondence, which assigns to every permutation of n letters a pair of tableaux of the same shape. One use of the RSK correspondence is to find the length of the longest increas-

1	2	4
3	6	
5		

Figure 2 A Young tableau.

ing subsequence in the vector of values of a given permutation. It turns out that this length is the same as the length of the first row of either of the tableaux to which the permutation corresponds under the RSK mapping. This fact gives us a good way, algorithmically speaking, of finding the length of the longest increasing subsequence of a given permutation.

Now suppose that $u_k(n)$ is the number of permutations of n letters that have no increasing subsequence of length greater than k . A spectacular theorem of Gessel (1990) states that

$$\sum_{n \geq 0} \frac{u_k(n)}{n!^2} x^{2n} = \det(I_{|i-j|}(2x))_{i,j=1,\dots,k}, \quad (5)$$

in which $I_\nu(t)$ is (the modified Bessel function)

$$I_\nu(t) = \sum_{j=0}^{\infty} \frac{(\frac{1}{2}t)^{2j+\nu}}{j!(j+\nu)!}.$$

At any rate, it seems fairly “spectacular” to me that when you place various infinite series such as the above into a $k \times k$ determinant and then expand the determinant, you should find that the coefficient of x^{2n} , when multiplied by $n!^2$, is exactly the number of permutations of n letters with no increasing subsequence longer than k .

Let’s evaluate one of these determinants, say the one with $k = 2$. We find that

$$\det(I_{|i-j|}(2x))_{i,j=1,2} = I_0^2 - I_1^2,$$

which of course factors as $(I_0 + I_1)(I_0 - I_1)$. The arguments of the I_ν are all $2x$ and have been omitted.

When $k = 3$, no such factorization occurs. If you ask your CAS for this determinant when $k = 4$, it will show you

$$\begin{aligned} & I_0^4 - 3I_0^2I_1^2 + I_1^4 + 4I_0I_1^2I_2 \\ & - 2I_0^2I_2^2 - 2I_1^2I_2^2 + I_2^4 - 2I_1^3I_3 \\ & + 4I_0I_1I_2I_3 - 2I_1I_2^2I_3 - I_0^2I_3^2 + I_1^2I_3^2, \end{aligned}$$

where now we have abbreviated $I_\nu(2x)$ simply by I_ν . If we ask our CAS to factor this last expression, it

(surprisingly) replies with

$$\begin{aligned} & (I_0^2 - I_0I_1 - I_1^2 + 2I_1I_2 - I_2^2 - I_0I_3 + I_1I_3) \\ & \times (I_0^2 + I_0I_1 - I_1^2 - 2I_1I_2 - I_2^2 + I_0I_3 + I_1I_3), \end{aligned}$$

which is actually of the form $(A+B)(A-B)$, as a quick inspection will reveal.

We have now observed, experimentally, that for $k = 2$ and $k = 4$ Gessel’s $k \times k$ determinant has a nontrivial factorization of the form $(A+B)(A-B)$, in which A and B are certain polynomials of degree $k/2$ in the Bessel functions. Such a factorization of a large expression in terms of formal Bessel functions simply cannot be ignored. It demands explanation. Does this factorization extend to all even values of k ? It does. Can we say anything in general about what the factors mean? We can.

The key point, as it turns out, is that in Gessel’s determinant (5), the matrix entries depend only on $|i-j|$ (such a matrix is called a *Toeplitz* matrix). The determinants of such matrices have a natural factorization, as follows. If a_0, a_1, \dots is some sequence, and $a_{-i} = a_i$, then we have

$$\begin{aligned} & \det(a_{i-j})_{i,j=1}^{2m} \\ & = \det(a_{i-j} + a_{i+j-1})_{i,j=1}^m \det(a_{i-j} - a_{i+j-1})_{i,j=1}^m. \end{aligned}$$

When we apply this fact to the present situation it correctly reproduces the above factorizations for $k = 2, 4$, and generalizes them to all even k , as follows.

Let $y_k(n)$ be the number of Young tableaux of n cells whose first row is of length at most k , and let

$$U_k(x) = \sum_{n \geq 0} \frac{u_k(n)}{n!^2} x^{2n} \quad \text{and} \quad Y_k(x) = \sum_{n \geq 0} \frac{y_k(n)}{n!} x^n.$$

In terms of these two generating functions, the general factorization theorem states that

$$U_k(x) = Y_k(x)Y_k(-x) \quad (k = 2, 4, 6, \dots).$$

Why is it useful to have such factorizations? For one thing we can equate the coefficients of like powers of x on both sides of this factorization (try it!). We then find an interesting explicit formula that relates the number of Young tableaux of n cells whose first row is of length at most k , on the one hand, and the number of permutations of n letters that have no increasing subsequence of length greater than k , on the other. No more direct proof of this relationship is known. For more details and some further consequences, see Wilf (1992).

Moral: cherchez les factorisations!

Terri: I can confirm that ‘Gessel’ and ‘Bessel’ are two different mathematicians.

5 A Score for Sloane's Database

Here is a case study in which, as it happens, not only was Sloane's database utilized, but Sloane himself was one of the authors of the ensuing research paper.

Eric Weisstein, the creator of the invaluable Web resource *MathWorld*, became interested in the enumeration of 0-1 matrices whose eigenvalues are all positive real numbers. If $f(n)$ is the number of $n \times n$ matrices whose entries are all 0s and 1s and whose eigenvalues are all real and positive, then by computation, Weisstein found for $f(n)$ the values

$$1, 3, 25, 543, 29281 \quad (\text{for } n = 1, 2, \dots, 5).$$

Upon looking up this sequence in Sloane's database, Weisstein found, interestingly, that this sequence is identical, as far as it goes, with sequence A003024 in the database. The latter sequence counts vertex-labeled acyclic directed graphs ("digraphs") of n vertices, and so Weisstein's conjecture was born:

[T]he number of vertex-labeled acyclic digraphs of n vertices is equal to the number of $n \times n$ 0-1 matrices whose eigenvalues are all real and positive.

This conjecture was proved in McKay et al. (2003). En route to the proof of the result, the following somewhat surprising fact was shown.

Theorem 1. *If a 0-1 matrix A has only real positive eigenvalues, then those eigenvalues are all equal to 1.*

To prove this, let $\{\lambda_i\}_{i=1}^n$ be the eigenvalues of A . Then

$$\begin{aligned} 1 &\geq \frac{1}{n} \text{trace}(A) \quad (\text{since all } A_{i,i} \leq 1) \\ &= \frac{1}{n} (\lambda_1 + \lambda_2 + \dots + \lambda_n) \\ &\geq (\lambda_1 \lambda_2 \dots \lambda_n)^{1/n} \\ &= (\det A)^{1/n} \\ &\geq 1, \end{aligned}$$

in which the third line uses the arithmetic-geometric mean inequality, and the last line uses the fact that $\det A$ is a positive integer. Since the arithmetic and geometric means of the eigenvalues are equal, the eigenvalues are all equal, and in fact all $\lambda_i(A) = 1$.

The proof of the conjecture itself works by finding an explicit bijection between the two sets that are being counted. Indeed, let A be an $n \times n$ matrix of 0s and 1s with positive eigenvalues only. Then those eigenvalues are all 1s, so the diagonal of A is all 1s, whence the matrix $A - I$ also has solely 0s and 1s as its entries.

Regard $A - I$ as the vertex adjacency matrix of a digraph G . Then (it turns out that) G is acyclic.

Conversely, if G is such a digraph, let B be its vertex adjacency matrix. By renumbering the vertices of G , if necessary, B can be brought to triangular form with zero diagonal. Then $A = I + B$ is a 0-1 matrix with positive real eigenvalues only. But then the same must have been true for the matrix $I + B$ before simultaneously renumbering its rows and columns. For more details and more corollaries, see McKay et al. (2003).

Moral: look for your sequence in the online encyclopedia!

6 The Twenty-One-Stage Rocket

Now we'll describe a successful attack that was carried out by Andrews (1998) on the problem of evaluating the Mills-Robbins-Rumsey determinant, which is the determinant of the $n \times n$ matrix

$$M_n(\mu) = \left(\binom{i+j+\mu}{2j-i} \right)_{0 \leq i, j \leq n-1}. \quad (6)$$

This problem arose (Mills et al. 1987) in connection with the study of *plane partitions*. A plane partition of an integer n is an (infinite) array $n_{i,j}$ of nonnegative integers whose sum is n , subject to the restriction that the entries $n_{i,j}$ are nonincreasing across each row, and also down each column.

It turns out that $\det M_n(\mu)$ can be expressed neatly as a product, namely as

$$\det M_n(\mu) = 2^{-n} \prod_{k=0}^{n-1} \Delta_k(2\mu), \quad (7)$$

in which

$$\Delta_{2j}(\mu) = \frac{(\mu + 2j + 2)_j (\frac{1}{2}\mu + 2j + \frac{3}{2})_{j-1}}{(j)_j (\frac{1}{2}\mu + j + \frac{3}{2})_{j-1}},$$

and $(x)_j$ is the rising factorial $x(x+1) \dots (x+j-1)$.

The strategy of Andrews's proof is elegant in conception and difficult in execution: we are going to find an upper triangular matrix $E_n(\mu)$, whose diagonal entries are all 1s, such that

$$M_n(\mu) E_n(\mu) = L_n(\mu) \quad (8)$$

is a lower triangular matrix, with the numbers $\{\frac{1}{2}\Delta_{2j}(2\mu)\}_{j=0}^{n-1}$ on its diagonal. Of course, if we can do this, then from (8), since $\det E_n(\mu) = 1$, we will have proved the theorem (7), since the determinant of the product of two matrices is the product of their determinants, and the determinant of a triangular matrix (i.e., of a matrix all of whose entries below the diagonal are 0s) is simply the product of its diagonal entries.

But how shall we find this matrix $E_n(\mu)$? By holding tightly to the hand of our computer and letting it guide us there. More precisely,

- (i) we will look at the matrix $E_n(\mu)$ for various small values of n , and from those data we will conjecture the formula for the general (i, j) entry of the matrix; and then
- (ii) we will (well actually “we” won’t, but Andrews did) prove that the conjectured entries of the matrix are correct.

It was in step (ii) above that an extraordinary twenty-one-stage event occurred which was successfully managed by Andrews. What he did was to set up a system of twenty-one propositions, each of them a fairly technical hypergeometric identity. Next, he carried out a simultaneous induction on these twenty-one propositions. That is to say, he showed that if, say, the thirteenth proposition was true for a certain value of n , then so was the fourteenth, etc., and if they were all true for that value of n , then the first proposition was true for $n + 1$. The reader should be sure to look at Andrews (1998) to gain more of the flavor and substance of what was done than can be conveyed in this short summary.

Here we will confine ourselves to a few comments about step (i) of the program above. So, let’s look at the matrix $E_n(\mu)$ for some small values of n . The condition that $E_n(\mu)$ is upper triangular with 1s on the diagonal means that

$$\sum_{k=0}^{j-1} (M_n)_{i,k} e_{k,j} = -(M_n)_{i,j},$$

for $0 \leq i \leq j-1$ and $1 \leq j \leq n-1$. We can regard these as $\binom{n}{2}$ equations in the $\binom{n}{2}$ above-diagonal entries of $E_n(\mu)$ and we can ask our CAS to find those entries, for some small values of n . Here is $E_4(\mu)$:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -\frac{1}{\mu+2} & \frac{6(\mu+5)}{(\mu+2)(\mu+3)(2\mu+11)} \\ 0 & 0 & 1 & -\frac{6(\mu+5)}{(\mu+3)(2\mu+11)} \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

At this point the news is all good. While it is true that the matrix entries are fairly complicated, the fact that leaps off the page and warms the heart of the experimental mathematician is that all of the polynomials in μ factor into linear factors with pleasant-looking integer coefficients. So there is hope for conjecturing a general form of the E matrix. Will this benign situation

persist when $n = 5$? A further computation reveals that $E_5(\mu)$ is as shown in figure 3. Now it is a “certainty” that some nice formulas exist for the entries of the general matrix $E_n(\mu)$. The Rate package, described in section 2.3, would certainly facilitate the next step, which is to find general formulas for the entries of the E matrix. The final result is that the (i, j) entry of $E_n(\mu)$ is 0 if $i > j$ and

$$\frac{(-1)^{j-i} (i)_{2(j-i)} (2\mu + 2j + i + 2)_{j-i}}{4^{j-i} (j-i)! (\mu + i + 1)_{j-i} (\mu + j + i + \frac{1}{2})_{j-i}}$$

otherwise.

After divining that the E matrix has the above form, Andrews now faced the task of proving that it works, i.e., that $M_n E_n(\mu)$ is lower triangular and has the diagonal entries specified above. It was in this part of the work that the twenty-one-fold induction was unleashed. Another proof of the evaluation of the Mills–Robbins–Rumsey determinant is in Petkovšek and Wilf (1996). That proof begins with Andrews’s discovery of the above form of the $E_n(\mu)$ matrix, and then uses the machinery of the so-called WZ method (Petkovšek et al. 1996), instead of a twenty-one-stage induction, to prove that the matrix performs the desired triangulation (8).

Moral: never give up, even when defeat seems certain.

7 The Computation of π

In 1997, a remarkable formula for π was found (Bailey et al. 1997). This formula permits the computation of just a single hexadecimal digit of π , if desired, using minimal space and time. For example, we might compute the trillionth digit of π , without ever having to deal with any of the earlier ones, in a time that is faster than what we might attain if we had to calculate all of the first trillion digits. For example, Bailey et al. found that in the hexadecimal expansion of π , the block of fourteen digits in positions 10^{10} through $10^{10} + 13$ are 921C73C6838FB2. The formula is

$$\pi = \sum_{i=0}^{\infty} \frac{1}{16^i} \left(\frac{4}{8i+1} - \frac{2}{8i+4} - \frac{1}{8i+5} - \frac{1}{8i+6} \right). \quad (9)$$

Terri: I can confirm that the letters are OK here – this is a hexadecimal number.

In our discussion here we will limit ourselves to describing how we might have found the specific expansion (9) once we had decided that an interesting expansion of the form

$$\pi = \sum_{i=0}^{\infty} \frac{1}{c^i} \sum_{k=1}^{b-1} \frac{a_k}{bi+k}. \quad (10)$$

might exist. This, of course, leaves open the question of how the discovery of the form (10) was singled out in the first place.

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -\frac{1}{\mu+2} & \frac{6(\mu+5)}{(\mu+2)(\mu+3)(2\mu+11)} & -\frac{30(\mu+6)}{(\mu+2)(\mu+3)(\mu+4)(2\mu+15)} \\ 0 & 0 & 1 & -\frac{6(\mu+5)}{(\mu+3)(2\mu+11)} & \frac{30(\mu+6)}{(\mu+3)(\mu+4)(2\mu+15)} \\ 0 & 0 & 0 & 1 & -\frac{6(2\mu+13)}{(\mu+4)(2\mu+15)} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Figure 3 The upper triangular matrix $E_5(\mu)$.

The strategy will be to use the linear dependency algorithm, described above in section 2.4. More precisely, we want to find a nontrivial integer linear combination of π and the seven numbers

$$\alpha_k = \sum_{i=0}^{\infty} \frac{1}{(8i+k)16^i} \quad (k = 1, \dots, 7)$$

that sums to 0. As in equation (3), we now compute the seven numbers α_j and we look for a relation

$$m\pi + m_1\alpha_1 + m_2\alpha_2 + \dots + m_7\alpha_7 = 0 \quad (m, m_i \in \mathbb{Z})$$

using, for example, the Maple `IntegerRelations` package. The output vector,

$$(m, m_1, m_2, \dots, m_7) = (1, -4, 0, 0, 2, 1, 1, 0),$$

yields the identity (9). You should do this calculation for yourself, then prove that the apparent identity is in fact true, and, finally, look for something similar that uses powers of 64 instead of 16. Good luck!

Moral: even as late as the year 1997 C.E., something new and interesting was said about the number π .

8 Conclusions

When computers first appeared in mathematicians' environments the almost universal reaction was that they would never be useful for proving theorems since a computer can never investigate infinitely many cases, no matter how fast it is. But computers are useful for proving theorems despite that handicap. We have seen several examples of how a mathematician can act in concert with a computer to explore a world within mathematics. From such explorations there can grow understanding, and conjectures, and roads to proofs, and phenomena that would not have been imaginable in the pre-computer era. This role of computation within pure mathematics seems destined only to expand over the coming years and to be imbued into our students

along with EUCLID's [VI.2] axioms and other staples of mathematical education.

At the other end of the rainbow there may lie a more far-reaching role for computers. Perhaps one day we will be able to input some hypotheses and a desired conclusion, press the "Enter" key, and get a printout of a proof. There are a few fields of mathematics in which we can do such things, notably in the proofs of identities (Petkovšek et al. 1996; Greene and Wilf 2007), but in general the road to that brave new world remains long and uncharted.

Further Reading

- Andrews, G. E. 1998. Pfaff's method. I. The Mills-Robbins-Rumsey determinant. *Discrete Mathematics* 193:43-60.
- Bailey, D. H., and S. Plouffe. 1997. Recognizing numerical constants. In *Proceedings of the Organic Mathematics Workshop, 12-14 December 1995, Simon Fraser University*. Conference Proceedings of the Canadian Mathematical Society, volume 20. Ottawa: Canadian Mathematical Society.
- Bailey, D. H., P. Borwein, and S. Plouffe. 1997. On the rapid computation of various polylogarithmic constants. *Mathematics of Computation* 66:903-13.
- Borwein, J., and D. H. Bailey. 2003. *Mathematics by Experiment: Plausible Reasoning in the 21st Century*. Wellesley, MA: A. K. Peters.
- Borwein, J., D. H. Bailey, and R. Girgensohn. 2004. *Experimentation in Mathematics: Computational Paths to Discovery*. Wellesley, MA: A. K. Peters.
- Calkin, N., and H. S. Wilf. 2000. Recounting the rationals. *American Mathematical Monthly* 107:360-63.
- Ferguson, H.R.P., and R. W. Forcade. 1979. Generalization of the Euclidean algorithm for real numbers to all dimensions higher than two. *Bulletin of the American Mathematical Society* 1:912-14.
- Gessel, I. 1990. Symmetric functions and P -recursiveness. *Journal of Combinatorial Theory A* 53:257-85.
- Graham, R. L., D. E. Knuth, and O. Patashnik. 1989. *Concrete Mathematics*. Reading, MA: Addison-Wesley.

- Greene, C., and Wilf, H. S. 2007. Closed form summation of C -finite sequences. *Transactions of the American Mathematical Society* 359:1161–89.
- Lenstra, A. K., H. W. Lenstra Jr., and L. Lovász. 1982. Factoring polynomials with rational coefficients. *Mathematische Annalen* 261(4):515–34.
- McKay, B. D., F. E. Oggier, G. F. Royle, N.J.A. Sloane, I. M. Wanless, and H. S. Wilf. 2004. Acyclic digraphs and eigenvalues of $(0, 1)$ -matrices. *Journal of Integer Sequences* 7: 04.3.3.
- Mills, W. H., D. P. Robbins, and H. Rumsey Jr. 1987. Enumeration of a symmetry class of plane partitions. *Discrete Mathematics* 67:43–55.
- Petkovšek, M., and H. S. Wilf. 1996. A high-tech proof of the Mills–Robbins–Rumsey determinant formula. *Electronic Journal of Combinatorics* 3:R19.
- Petkovšek, M., H. S. Wilf, and D. Zeilberger. 1996. $A = B$. Wellesley, MA: A. K. Peters.
- Wilf, H. S. 1992. Ascending subsequences and the shapes of Young tableaux. *Journal of Combinatorial Theory A* 60: 155–57.
- . 1994. *generatingfunctionology*, 2nd edn. New York: Academic Press. (This can also be downloaded at no charge from the author's Web site.)

VIII.6 Advice to a Young Mathematician

The most important thing that a young mathematician needs to learn is of course mathematics. However, it can also be very valuable to learn from the experiences of other mathematicians. The five contributors to this article were asked to draw on their experiences of mathematical life and research, and to offer advice that they might have liked to receive when they were just setting out on their careers. (The title of this entry is a nod to Sir Peter Medawar's well-known book, *Advice to a Young Scientist*.) The resulting contributions were every bit as interesting as we had expected; what was more surprising was that there was remarkably little overlap between the contributions. So here they are, five gems intended for young mathematicians but surely destined to be read and enjoyed by mathematicians of all ages.

I. Sir Michael Atiyah

Warning

What follows is very much a personal view based on my own experience and reflecting my personality, the type of mathematics that I work on, and my style of work. However, mathematicians vary widely in all these characteristics and you should follow your own instinct.

You may learn from others but interpret what you learn in your own way. Originality comes by breaking away, in some respects, from the practice of the past.

Motivation

A research mathematician, like a creative artist, has to be passionately interested in the subject and fully dedicated to it. Without strong internal motivation you cannot succeed, but if you enjoy mathematics the satisfaction you can get from solving hard problems is immense.

The first year or two of research is the most difficult. There is so much to learn. One struggles unsuccessfully with small problems and one has serious doubts about one's ability to prove anything interesting. I went through such a period in my second year of research, and Jean-Pierre Serre, perhaps the outstanding mathematician of my generation, told me that he too had contemplated giving up at one stage.

Only the mediocre are supremely confident of their ability. The better you are, the higher the standards you set yourself—you can see beyond your immediate reach.

Terri: Tim prefers 'set yourself' to 'OK for yourself' – OK to keep it as it is?

Many would-be mathematicians also have talents and interests in other directions and they may have a difficult choice to make between embarking on a mathematical career and pursuing something else. The great Gauss is reputed to have wavered between mathematics and philology, Pascal deserted mathematics at an early age for theology, while Descartes and Leibniz are also famous as philosophers. Some mathematicians move into physics (e.g., Freeman Dyson) while others (e.g., Harish Chandra, Raoul Bott) have moved the other way. You should not regard mathematics as a closed world, and the interaction between mathematics and other disciplines is healthy both for the individual and for society.

Psychology

Because of the intense mental concentration required in mathematics, psychological pressures can be considerable, even when things are going well. Depending on your personality this may be a major or only a minor problem, but one can take steps to reduce the tension. Interaction with fellow students—attending lectures, seminars, and conferences—both widens one's horizons and provides important social support. Too much isolation and introspection can be dangerous,

and time spent in apparently idle conversation is not really wasted.

Collaboration, initially with fellow students or one's supervisor, has many benefits, and long-term collaboration with coworkers can be extremely fruitful both in mathematical terms and at the personal level. There is always the need for hard quiet thought on one's own, but this can be enhanced and balanced by discussion and exchange of ideas with friends.

Problems versus Theory

Mathematicians are sometimes categorized as either "problem solvers" or "theorists." It is certainly true that there are extreme cases that highlight this division (Erdős versus Grothendieck, for example) but most mathematicians lie somewhere in between, with their work involving both the solution of problems and the development of some theory. In fact, a theory that does not lead to the solution of concrete and interesting problems is not worth having. Conversely, any really deep problem tends to stimulate the development of theory for its solution (Fermat's last theorem being a classic example).

What bearing does this have on a beginning student? Although one has to read books and papers and absorb general concepts and techniques (theory), realistically, a student has to focus on one or more specific problems. This provides something to chew on and to test one's mettle. A definite problem, which one struggles with and understands in detail, is also an invaluable benchmark against which to measure the utility and strength of available theories.

Depending on how the research goes, the eventual Ph.D. thesis may strip away most of the theory and focus only on the essential problem, or else it may describe a wider scenario into which the problem naturally fits.

The Role of Curiosity

The driving force in research is curiosity. When is a particular result true? Is that the best proof, or is there a more natural or elegant one? What is the most general context in which the result holds?

If you keep asking yourself such questions when reading a paper or listening to a lecture, then sooner or later a glimmer of an answer will emerge—some possible route to investigate. When this happens to me I always take time out to pursue the idea to see where it leads or whether it will stand up to scrutiny. Nine

times out of ten it turns out to be a blind alley, but occasionally one strikes gold. The difficulty is in knowing when an idea that is initially promising is in fact going nowhere. At this stage one has to cut one's losses and return to the main road. Often the decision is not clear-cut, and in fact I frequently return to a previously discarded idea and give it another try.

Ironically, good ideas can emerge unexpectedly from a bad lecture or seminar. I often find myself listening to a lecture where the result is beautiful and the proof ugly and complicated. Instead of trying to follow a messy proof on the blackboard, I spend the rest of the hour thinking about producing a more elegant proof. Usually, but not always, without success, but even then my time is better spent, since I have thought hard about the problem in my own way. This is much better than passively following another person's reasoning.

Examples

If you are, like me, someone who prefers large vistas and powerful theories (I was influenced but not converted by Grothendieck), then it is essential to be able to test general results by applying them to simple examples. Over the years I have built up a large array of such examples, drawn from a variety of fields. These are examples where one can do concrete calculations, sometimes with elaborate formulas, that help to make the general theory understandable. They keep your feet on the ground. Interestingly enough, Grothendieck eschewed examples, but fortunately he was in close touch with Serre, who was able to rectify this omission. There is no clear-cut distinction between example and theory. Many of my favorite examples come from my early training in classical projective geometry: the twisted cubic, the quadric surface, or the Klein representation of lines in 3-space. Nothing could be more concrete or classical and all can be looked at algebraically or geometrically, but each illustrates and is the first case in a large class of examples which then become a theory: the theory of rational curves, of homogeneous spaces, or of Grassmannians.

Another aspect of examples is that they can lead off in different directions. One example can be generalized in several different ways or illustrate several different principles. For instance, the classical conic is a rational curve, a quadric, and a Grassmannian all in one.

But most of all a good example is a thing of beauty. It shines and convinces. It gives insight and understanding. It provides the bedrock of belief.

Proof

We are all taught that “proof” is the central feature of mathematics, and Euclidean geometry with its careful array of axioms and propositions has provided the essential framework for modern thought since the Renaissance. Mathematicians pride themselves on absolute certainty, in comparison with the tentative steps of natural scientists, let alone the woolly thinking of other areas.

It is true that, since Gödel, absolute certainty has been undermined, and the more mundane assault of computer proofs of interminable length has induced some humility. Despite all this, proof retains its cardinal role in mathematics, and a serious gap in your argument will lead to your paper being rejected.

However, it is a mistake to identify research in mathematics with the process of producing proofs. In fact, one could say that all the really creative aspects of mathematical research precede the proof stage. To take the metaphor of the “stage” further, you have to start with the idea, develop the plot, write the dialogue, and provide the theatrical instructions. The actual production can be viewed as the “proof”: the implementation of an idea.

In mathematics, ideas and concepts come first, then come questions and problems. At this stage the search for solutions begins, one looks for a method or strategy. Once you have convinced yourself that the problem has been well-posed, and that you have the right tools for the job, you then begin to think hard about the technicalities of the proof.

Before long you may realize, perhaps by finding counterexamples, that the problem was incorrectly formulated. Sometimes there is a gap between the initial intuitive idea and its formalization. You left out some hidden assumption, you overlooked some technical detail, you tried to be too general. You then have to go back and refine your formalization of the problem. It would be an unfair exaggeration to say that mathematicians rig their questions so that they can answer them, but there is undoubtedly a grain of truth in the statement. The art in good mathematics, and mathematics is an art, is to identify and tackle problems that are both interesting and solvable.

Proof is the end product of a long interaction between creative imagination and critical reasoning. Without proof the program remains incomplete, but without the imaginative input it never gets started. One can see here an analogy with the work of the creative artist in

other fields: writer, painter, composer, or architect. The vision comes first, it develops into an idea that gets tentatively sketched out, and finally comes the long technical process of erecting the work of art. But the technique and the vision have to remain in touch, each modifying the other according to its own rules.

Strategy

In the previous section I discussed the philosophy of proof and its role in the whole creative process. Now let me turn to the most down-to-earth question of interest to the young practitioner. What strategy should one adopt? How do you actually go about finding a proof?

This question makes little sense in the abstract. As I explained in the previous section a good problem always has antecedents: it arises from some background, it has roots. You have to understand these roots in order to make progress. That is why it is always better to find your own problem, asking your own questions, rather than getting it on a plate from your supervisor. If you know where a problem comes from, why the question has been asked, then you are halfway toward its solution. In fact, asking the right question is often as difficult as solving it. Finding the right context is an essential first step.

So, in brief, you need to have a good knowledge of the history of the problem. You should know what sort of methods have worked with similar problems and what their limitations are.

It is a good idea to start thinking hard about a problem as soon as you have fully absorbed it. To get to grips with it, there is no substitute for a hands-on approach. You should investigate special cases and try to identify where the essential difficulty lies. The more you know about the background and previous methods, the more techniques and tricks you can try. On the other hand, ignorance is sometimes bliss. J. E. Littlewood is reported to have set each of his research students to work on a disguised version of the Riemann hypothesis, letting them know what he had done only after six months. He argued that the student would not have the confidence to attack such a famous problem directly, but might make progress if not told of the fame of his opponent! The policy may not have led to a proof of the Riemann hypothesis, but it certainly led to resilient and battle-hardened students.

My own approach has been to try to avoid the direct onslaught and look for indirect approaches. This involves connecting your problem with ideas and techniques from different fields that may shed unexpected

light on it. If this strategy succeeds, it can lead to a beautiful and simple proof, which also “explains” why something is true. In fact, I believe the search for an explanation, for understanding, is what we should really be aiming for. Proof is simply part of that process, and sometimes its consequence.

As part of the search for new methods it is a good idea to broaden your horizons. Talking to people will extend your general education and will sometimes introduce you to new ideas and techniques. Very occasionally you may get a productive idea for your own research or even for a new direction.

If you need to learn a new subject, consult the literature but, even better, find a friendly expert and get instruction “from the horse’s mouth”—it gives more insight more quickly.

As well as looking forward, and being alert to new developments, you should not forget the past. Many powerful mathematical results from earlier eras have got buried and have been forgotten, coming to light only when they have been independently rediscovered. These results are not easy to find, partly because terminology and style change, but they can be gold mines. As usual with gold mines, you have to be lucky to strike one, and the rewards go to the pioneers.

Independence

At the start of your research your relationship with your supervisor can be crucial, so choose carefully, bearing in mind subject matter, personality, and track record. Few supervisors score highly on all three. Moreover, if things do not work out well during the first year or so, or if your interests diverge significantly, then do not hesitate to change supervisors or even universities. Your supervisor will not be offended and may even be relieved!

Sometimes you may be part of a large group and may interact with other members of the faculty, so that you effectively have more than one supervisor. This can be helpful in that it provides different inputs and alternative modes of work. You may also learn much from fellow students in such large groups, which is why choosing a department with a large graduate school is a good idea.

Once you have successfully earned your Ph.D. you enter a new stage. Although you may still carry on collaborating with your supervisor and remain part of the same research group, it is healthy for your future development to move elsewhere for a year or more. This

opens you up to new influences and opportunities. This is the time when you have the chance to carve out a niche for yourself in the mathematical world. In general, it is not a good idea to continue too closely in the line of your Ph.D. thesis for too long. You have to show your independence by branching out. It need not be a radical change of direction but there should be some clear novelty and not simply a routine continuation of your thesis.

Style

In writing up your thesis your supervisor will normally assist you in the manner of presentation and organization. But acquiring a personal style is an important part of your mathematical development. Although the needs may vary, depending on the kind of mathematics, many aspects are common to all subjects. Here are a number of hints on how to write a good paper.

- (i) Think through the whole logical structure of the paper before you start to write.
- (ii) Break up long complex proofs into short intermediate steps (lemmas, propositions, etc.) that will help the reader.
- (iii) Write clear coherent English (or the language of your choice). Remember that mathematics is also a form of literature.
- (iv) Be as succinct as it is possible to be while remaining clear and easy to understand. This is a difficult balance to achieve.
- (v) Identify papers that you have enjoyed reading and imitate their style.
- (vi) When you have finished writing the bulk of your paper go back and write an introduction that explains clearly the structure and main results as well as the general context. Avoid unnecessary jargon and aim at a general mathematical reader, not just a narrow expert.
- (vii) Try out your first draft on a colleague and take heed of any suggestions or criticisms. If even your close friend or collaborator has difficulty understanding it, then you have failed and need to try harder.
- (viii) If you are not in a desperate hurry to publish, put your paper aside for a few weeks and work on something else. Then return to your paper and read it with a fresh mind. It will read differently and you may see how to improve it.
- (ix) Do not hesitate to rewrite the paper, perhaps from a totally new angle, if you become convinced

Terri: Tim would prefer to keep this as we've been more careful to keep the authors' voices in these articles than in most of the Companion. OK?

that this will make it clearer and easier to read. Well-written papers become “classics” and are widely read by future mathematicians. Badly written papers are ignored or, if they are sufficiently important, they get rewritten by others.

II. Béla Bollobás

“There is no permanent place in this world for ugly mathematics,” wrote Hardy; I believe that it is just as true that there is no place in this world for unenthusiastic, dour mathematicians. Do mathematics only if you are passionate about it, only if you would do it even if you had to find the time for it after a full day’s work in another job. Like poetry and music, mathematics is not an occupation but a vocation.

Taste is above everything. It is a miracle of our subject that there seems to be a consensus as to what constitutes good mathematics. You should work in areas that are important and unlikely to dry up for a long time, and you should work on problems that are beautiful and important: in a good area there will be plenty of these, and not just a handful of well-known problems. Indeed, aiming too high all the time may lead to long barren periods: these may be tolerated at some stage of your life, but at the beginning of your career it is best to avoid them.

Strive for a *balance* in your mathematical activity: research should and does come first for real mathematicians, but in addition to doing research, do plenty of reading and teach well. Have fun with mathematics at all levels, even if it has (almost) no bearing on your research. Teaching should not be a burden but a source of inspiration.

Research should never be a chore (unlike writing up): you should choose problems that you find it difficult *not* to think about. This is why it is good if you get *yourself* hooked on problems rather than working on problems as if you were doing a task imposed on you. At the very beginning of your career, when you are a research student, you should use your experienced supervisor to help you judge problems that you have found and like, rather than working on a problem that he has handed to you, which may not be to your taste. After all, your supervisor should have a fairly good idea whether a certain problem is worth your efforts or not, while he may not yet know your strength and taste. Later in your career, when you can no longer rely on your supervisor, it is frequently inspiring to talk to sympathetic colleagues.

I would recommend that at any one time you have problems of two types to work on.

- (i) A “dream”: a big problem that you would love to solve, but you cannot reasonably *expect* to solve.
- (ii) Some very worthwhile problems that you feel you should have a good chance of solving, given enough time, effort, and luck.

In addition, there are two more types you should consider, although these are less important than the previous ones.

- (i) From time to time, work on problems that should be below your dignity and that you can be confident of doing rather quickly, so that time spent on them will not jeopardize your success with the proper problems.
- (ii) On an even lower level, it is always fun to do problems that are not really research problems (although they may have been some years ago) but are beautiful enough to spend time on: doing them will give you pleasure and will sharpen your ability to be inventive.

Be patient and persistent. When thinking about a problem, perhaps the most useful device you can employ is to bear the problem in mind all the time: it worked for Newton, and it has worked for many a mortal as well. Give yourself time, especially when attacking major problems; promise yourself that you will spend a certain amount of time on a big problem without expecting much, and after that take stock and decide what to do next. Give your approach a chance to work, but do not be so wrapped up in it that you miss other ways of attacking the problem. Be mentally agile: as Paul Erdős put it, keep your brain open.

Do not be afraid to make mistakes. A mistake for a chess player is fatal; for a mathematician it is par for the course. What you should be terrified of is a blank sheet in front of you after having thought about a problem for a little while. If after a session your wastepaper basket is full of notes of failed attempts, you may still be doing very well. Avoid pedestrian approaches, but always be happy to put in work. In particular, doing the simplest cases of a problem is unlikely to be a waste of time and may well turn out to be very useful.

When you spend a significant amount of time on a problem, it is easy to underestimate the progress you have made, and it is equally easy to overestimate your ability to remember it all. It is best to write down even

your very partial results: there is a good chance that your notes will save you a great deal of time later.

If you are lucky enough to have made a breakthrough, it is natural to feel fed up with the project and to want to rest on your laurels. Resist this temptation and see what else your breakthrough may give you.

As a young mathematician, your main advantage is that you have plenty of time for research. You may not realize it, but it is very unlikely that you will ever again have as much time as you do at the beginning of your career. Everybody feels that there is not enough time to do mathematics, but as the years pass this feeling gets more and more acute, and more and more justified.

Turning to *reading*, young people are at a disadvantage when it comes to the amount of mathematics they have read, so to compensate for this, read as much as you can, both in your general area and in mathematics as a whole. In your own research area, make sure that you read many papers written by the best people. These papers are often not as carefully written as they could be, but the quality of the ideas and results should amply reward you for the effort you have to make to read them. Whatever you read, be alert: try to anticipate what the author will do and try to think up a better attack. When the author takes the route you had in mind, you will be happy, and when he chooses to go a different way, you can look forward to finding out why. Ask yourself questions about the results and proofs, even if they seem simpleminded: they will greatly help your understanding.

On the other hand, it is often useful *not* to read up everything about an open problem you are about to attack: once you have thought deeply about it and apparently got nowhere, you can (and should) read the failed attempts of others.

Keep your ability to be surprised, do not take phenomena for granted, appreciate the results and ideas you read. It is all too easy to think that you know what is going on: after all, you have just read the proof. Outstanding people often spend a great deal of time digesting new ideas. It is not enough for them to know a circle of theorems and understand their proofs: they want to feel them in their blood.

As your career progresses, always keep your mind open to new ideas and new directions: the mathematical landscape changes all the time, and you will probably have to as well if you do not want to be left behind. Always sharpen your tools and learn new ones.

Above everything, *enjoy mathematics and be enthusiastic about it*. Enjoy your research, look forward to read-

ing about new results, feed the love of mathematics in others, and even in your recreation have fun with mathematics by thinking about beautiful little problems you come across or hear from your colleagues.

If I wanted to sum up the advice we should all follow in order to be successful in the sciences and the arts, I could hardly do better than recall what Vitruvius wrote over two thousand years ago:

Neque enim ingenium sine disciplina aut disciplina sine ingenio perfectum artificem potest efficere.

For neither genius without learning nor learning without genius can make a perfect artist.

III. Alain Connes

Mathematics is the backbone of modern science and a remarkably efficient source of new concepts and tools for understanding the “reality” in which we participate. The new concepts themselves are the result of a long process of “distillation” in the alembic of human thought.

I was asked to write some advice for young mathematicians. My first observation is that each mathematician is a special case, and in general mathematicians tend to behave like “fermions,” i.e., they avoid working in areas that are too trendy, whereas physicists behave a lot more like “bosons,” which coalesce in large packs, often “overselling” their achievements—an attitude that mathematicians despise.

It might be tempting at first to regard mathematics as a collection of separate branches, such as geometry, algebra, analysis, number theory, etc., where the first is dominated by the attempt to understand the concept of “space,” the second by the art of manipulating symbols, the third by access to “infinity” and the “continuum,” and so on.

This, however, does not do justice to one of the most important features of the mathematical world, namely that it is virtually impossible to isolate any of the above parts from the others without depriving them of their essence. In this way the corpus of mathematics resembles a biological entity, which can only survive as a whole and which would perish if separated into disjoint pieces.

The scientific life of mathematicians can be pictured as an exploration of the geography of the “mathematical reality” which they unveil gradually in their own private mental frame.

This process often begins with an act of rebellion against the dogmatic descriptions of that space that

can be found in existing books. Young, prospective mathematicians begin to realize that their own perception of the mathematical world captures some features that do not quite fit in with the existing dogma. This initial rebellion is, in most cases, due to ignorance, but it can nevertheless be beneficial, as it frees people from reverence for authority and allows them to rely on their intuition, provided that that intuition can be backed up by actual proofs. Once a mathematician truly gets to know, in an original and “personal” manner, some small part of the mathematical world, however esoteric it may look at first,¹ the journey can properly start. It is of course vital not to break the “fil d’Arianne” (“Ariadne’s thread”): that way one can constantly keep a fresh eye on whatever one encounters along the way, but one can also go back to the source if one ever begins to feel lost.

It is also vital to keep moving. Otherwise, one risks confining oneself to a relatively small area of extreme technical specialization, thereby limiting one’s perception of the mathematical world and of its huge, even bewildering, diversity.

The fundamental point in this respect is that, even though many mathematicians have spent their lives exploring different parts of that world, with different perspectives, they all agree on its contours and interconnections. Whatever the origin of one’s journey, one day, if one walks far enough, one is bound to stumble on a well-known town: for instance, elliptic functions, modular forms, or zeta functions. “All roads lead to Rome,” and the mathematical world is “connected.” Of course, this is not to say that all parts of mathematics look alike, and it is worth quoting what Grothendieck says (in *Récoltes et Semailles*) in comparing the landscape of analysis in which he first worked with that of algebraic geometry, in which he spent the rest of his mathematical life:

Je me rappelle encore de cette impression saisissante (toute subjective certes), comme si je quittais des steppes arides et revêches, pour me retrouver soudain dans une sorte de “pays promis” aux richesses luxuriantes, se multipliant à l’infini partout où il plait à la main de se poser, pour cueillir ou pour fouiller.²

1. My own starting point was the localization of roots of polynomials. Fortunately, I was invited at a very early age to attend a conference in Seattle, at which I was introduced to the roots of all my future work on factors.

2. Translation: “I still remember this strong impression (completely subjective of course), as if I was leaving dry and gloomy steppes and finding myself suddenly in a sort of ‘promised land’ of luxuriant richness, which spread out to infinity wherever one might wish to put out one’s hand to gather from it or delve about in it.”

Most mathematicians adopt a pragmatic attitude and see themselves as explorers of this “mathematical world” whose existence they do not have any wish to question, and whose structure they uncover by a mixture of intuition and a great deal of rational thought. The former is not so different from “poetical desire” (as emphasized by the French poet Paul Valéry), while the latter requires intense periods of concentration.

Each generation builds a mental picture that reflects their own understanding of this world. They construct mental tools that penetrate more and more deeply into it, so that they can explore aspects of it that were previously hidden.

Where things get really interesting is when unexpected bridges emerge between parts of the mathematical world that were remote from each other in the mental picture that had been developed by previous generations of mathematicians. When this happens, one gets the feeling that a sudden wind has blown away the fog that was hiding parts of a beautiful landscape. In my own work this type of great surprise has come mostly from the interaction with physics. The mathematical concepts that arise naturally in physics often turn out to be fundamental, as Hadamard pointed out. For him they exhibit

not this short lived novelty which can too often influence the mathematician left to his own devices, but the infinitely fecund novelty that springs from the nature of things.

I will end this article with some more “practical” advice. Note, though, that each mathematician is a “special case” and one should not take the advice too seriously.

Walks. One very sane exercise, when fighting with a very complicated problem (often involving computations), is to go for a long walk (no paper or pencil) and do the computation in one’s head, irrespective of whether one initially feels that “it is too complicated to be done like that.” Even if one does not succeed, it trains the live memory and sharpens one’s skills.

Lying down. Mathematicians usually have a hard time explaining to their partner that the times when they work with most intensity are when they are lying down in the dark on a sofa. Unfortunately, with e-mail and the invasion of computer screens in all mathematical institutions, the opportunity to isolate oneself and concentrate is becoming rarer, and all the more valuable.

Terri: Tim would prefer to keep this sentence as it is. OK?

Being brave. There are several phases in the process that leads to the discovery of new mathematics. While the checking phase is scary, but involves just rationality and concentration, the first, more creative, phase is of a totally different nature. In some sense, it requires a kind of protection of one's ignorance, since this also protects one from the billions of reasons there will always be for not looking at a problem that has already been unsuccessfully attacked by many other mathematicians.

Setbacks. Throughout their working lives, including at the very early stages, mathematicians will receive preprints from competitors and feel disrupted. The only suggestion I have here is to try to convert this feeling of frustration into an injection of positive energy for working harder. However, this is not always easy.

Grudging approbation. A colleague of mine once said, "We [mathematicians] work for the grudging approbation of a few friends." It is true that, since research work is of a rather solitary nature, we badly need that approbation in one way or another, but quite frankly one should not expect much. In fact, the only real judge is oneself. Nobody else is in as good a position to know what work was involved, and caring too much about the opinion of others is a waste of time: so far no theorem has been proved as the result of a vote. As Feynman put it, "Why do you care what other people think?"

IV. Dusa McDuff

I started my adult life in a very different situation from most of my contemporaries. Always brought up to think I would have an independent career, I had also received a great deal of encouragement from my family and school to do mathematics. Unusually, my girls' school had a wonderful mathematics teacher who showed me the beauty of Euclidean geometry and calculus. In contrast, I did not respect the science teachers, and since those at university were not much better I never really learned any physics.

Very successful within this limited sphere, I was highly motivated to be a research mathematician. While in some respects I had enormous self-confidence, in other ways I grew to feel very inadequate. One basic problem was that somehow I had absorbed the message that women are second rate as far as professional life is concerned and are therefore to be ignored. I had no female friends and did not really value my kind of

intelligence, thinking it boring and practical (female), and not truly creative (male). There were many ways of saying this: women keep the home fires burning while men go out into the world, women are muses not poets, women do not have the true soul needed to be a mathematician, etc. And there still are many ways of saying this. Recently an amusing letter circulated among my feminist friends: it listed various common and contradictory prejudices in different scientific fields, the message being that women are perceived to be incapable of whatever is most valued.

Another problem that became apparent a little later was that I had managed to write a successful Ph.D. thesis while learning very little mathematics. My thesis was in von Neumann algebras, a specialized topic that did not relate to anything with real meaning for me. I could see no way forward in that field, and yet I knew almost nothing else. When I arrived in Moscow in my last year of graduate study, Gel'fand gave me a paper to read on the cohomology of the Lie algebra of vector fields on a manifold, and I did not know what cohomology was, what a manifold was, what a vector field was, or what a Lie algebra was.

Terri: strange
apostrophe is
correct here.

Though this ignorance was partly the fault of an over-specialized educational system, it also resulted from my lack of contact with the wider world of mathematics. I had solved the problem of how to reconcile being a woman with being a mathematician by essentially leading two separate lives. My isolation increased upon my return from Moscow. Having switched fields from functional analysis to topology, I had little guidance, and I was too afraid of appearing ignorant to ask many questions. Also, I had a baby while I was a postdoc, and was therefore very busy coping with practical matters. At that stage, with no understanding of the process of doing mathematics, I was learning mostly by reading, unaware of the essential role played by formulating questions and trying out one's own, perhaps naive, ideas. I also had no understanding of how to build a career. Good things do not just happen: one has to apply for fellowships and jobs and keep an eye out for interesting conferences. It would certainly have helped to have had a mentor to suggest better ways of dealing with all these difficulties.

I probably most needed to learn how to ask good questions. As a student, one's job is not only to learn enough to be able to answer questions posed by others, but also to learn how to frame questions that might lead somewhere interesting. When studying something new

I often used to start in the middle, using some complicated theory already developed by others. But often one sees further by starting with the simplest questions and examples, because that makes it easier to understand the basic problem and then perhaps to find a new approach to it. For example, I have always liked working with Gromov's nonsqueezing theorem in symplectic geometry, which imposes restrictions on the ways a ball can be manipulated in a symplectic way. This very fundamental and geometric result somehow resonates for me, and so forms a solid basis from which to start exploring.

These days people are much more aware that mathematics is a communal endeavor: even the most brilliant idea gets meaning only from its relation to the whole. Once one has an understanding of the context, it is often very important and fruitful to work by oneself. However, while one is learning it is vital to interact with others.

There have been many successful attempts to facilitate such communication, by changing the structure of buildings, of conferences and meetings, of departmental programs, and also, less formally, of seminars and lectures. It is amazing how the atmosphere in a seminar changes when a senior mathematician, instead of going to sleep or looking bored, asks questions that clarify and open up the discussion for everyone there. Often people (both young and old) are intimidated into silence because they fear showing their ignorance, lack of imagination, or other fatal defect. But in the face of a subject as difficult and beautiful as mathematics, everyone has something to learn from others. Now there are many wonderful small conferences and workshops, organized so that it is easy to have discussions both about the details of specific theories and also about formulating new directions and questions.

The problem of how to reconcile being a woman and a mathematician is still of concern, although the idea that mathematics is intrinsically unfeminine is much less prevalent. I do not think that we women are as fully present in the world of mathematics as we could be, but there are enough of us that we can no longer be dismissed as exceptions. I have found meetings intended primarily for women to be unexpectedly worthwhile; the atmosphere is different when a lecture room is full of women discussing mathematics. Also, as is increasingly understood, the real question is how *any* young person can build a satisfying personal life while still managing to be a creative mathematician. Once people

start working on this in a serious way, we will have truly come a long way.

V. Peter Sarnak

I have guided quite a number of Ph.D. students over the years, which perhaps qualifies me to write as an experienced mentor. When advising a brilliant student (and I have been fortunate enough to have had my fair share of these) the interaction is a bit like telling *someone* to dig for gold in some general area and offering just a few vague suggestions. Once they move into action with their skill and talent they find diamonds instead (and of course, after the fact one cannot resist saying "I told you so"). In these cases, and in most others as well, the role of a senior mentor is more like that of a coach: one provides encouragement and makes sure that the person being mentored is working on interesting problems and is aware of the basic tools that are available. Over the years I have found myself repeating certain comments and suggestions that may have been found useful. Here is a list of some of them.

(i) When learning an area, one should combine reading modern treatments with a study of the original papers, especially papers by the masters of our subject. One of the troubles with recent accounts of certain topics is that they can become too slick. As each new author finds cleverer proofs or treatments of a theory, the treatment evolves toward the one that contains the "shortest proofs." Unfortunately, these are often in a form that causes the new student to ponder, "How did anyone think of this?" By going back to the original sources one can usually see the subject evolving naturally and understand how it has reached its modern form. (There will remain those unexpected and brilliant steps at which one can only marvel at the genius of the inventor, but there are far fewer of these than you might think.) As an example, I usually recommend reading Weyl's original papers on the representation theory of compact Lie groups and the derivation of his character formula, alongside one of the many modern treatments. Similarly, I recommend his book *The Concept of a Riemann Surface* to someone who knows complex analysis and wants to learn about the modern theory of Riemann surfaces, which is of central importance to many areas of mathematics. It is also instructive to study the collected works of superb mathematicians such as Weyl. Besides learning their theorems one uncovers how their minds work. There is almost always a natural line of thought that leads from one paper to

Terri: again Tim would like this sentence to stay as it is. OK?

the next and certain developments are then appreciated as inevitable. This can be very inspiring.

(ii) On the other hand, you should question dogma and “standard conjectures,” even if these have been made by brilliant people. Many standard conjectures are made on the basis of special cases that one understands. Beyond that, they are sometimes little more than wishful thinking: one just hopes that the general picture is not significantly different from the picture that the special cases suggest. There are a number of instances that I know of where someone set out to prove a result that was generally believed to be true and made no progress until they seriously questioned it. Having said that, I also find it a bit irritating when, for no particularly good reason, skepticism is thrown on certain special conjectures, such as the Riemann hypothesis, or on their provability. While as a scientist one should certainly adopt a critical attitude (especially toward some of the artificial objects that we mathematicians have invented), it is important psychologically that we have beliefs about our mathematical universe and about what is true and what is provable.

(iii) Do not confuse “elementary” with “easy”: a proof can certainly be elementary without being easy. In fact, there are many examples of theorems for which a little sophistication makes the proof easy to understand and brings out the underlying ideas, whereas an elementary treatment that avoids sophisticated notions hides what is going on. At the same time, beware of equating sophistication with quality or with the “beef of an argument” (an expression that I apparently like to use a lot in this context: many of my former students have teased me about it). There is a tendency among some young mathematicians to think that using fancy and sophisticated language means that what they are doing is deep. Nevertheless, modern tools are powerful when they are understood properly and when they are combined with new ideas. Those working in certain fields (number theory, for example) who do not put in the time and substantial effort needed to learn these tools are putting themselves at a great disadvantage. Not to learn the tools is like trying to demolish a building with just a chisel. Even if you are very adept at using the chisel, somebody with a bulldozer will have a huge advantage and will not need to be nearly as skilful as you.

(iv) Doing research in mathematics is frustrating and if being frustrated is something you cannot get used to, then mathematics may not be an ideal occupation for

you. Most of the time one is stuck, and if this is not the case for you, then either you are exceptionally talented or you are tackling problems that you knew how to solve before you started. There is room for some work of the latter kind, and it can be of a high quality, but most of the big breakthroughs are earned the hard way, with many false steps and long periods of little progress, or even negative progress. There are ways to make this aspect of research less unpleasant. Many people these days work jointly, which, besides the obvious advantage of bringing different expertise to bear on a problem, allows one to share the frustration. For most people this is a big positive (and in mathematics the corresponding sharing of the joy and credit on making a breakthrough has not, so far at least, led to many big fights in the way that it has in some other areas of science). I often advise students to try to have a range of problems at hand at any given moment. The least challenging should still be difficult enough that solving it will give you satisfaction (for without that, what is the point?) and with luck it will be of interest to others. Then you should have a range of more challenging problems, with the most difficult ones being central unsolved problems. One should attack these on and off over time, looking at them from different points of view. It is important to keep exposing oneself to the possibility of solving very difficult problems and perhaps benefiting from a bit of luck.

(v) Go to your departmental colloquium every week, and hope that its organizers have made some good choices for speakers. It is important to have a broad awareness of mathematics. Besides learning about interesting problems and progress that people are making in other fields, you can often have an idea stimulated in your mind when the speaker is talking about something quite different. Also, you may learn of a technique or theory that could be applied to one of the problems that you are working on. In recent times, a good number of the most striking resolutions of long-standing problems have come about from an unexpected combination of ideas from different areas of mathematics.

Terri: and again, Tim would prefer things to stay as they are here. OK?